

# The American Economic Review

COK-H01164-96-PO14422  
ARTICLES

397  
.001

ROBERT M. SOLOW

Growth Theory and After

JEFFREY A. FRANKEL AND KATHARINE ROCKETT

International Macroeconomic Policy Coordination When  
Policymakers Do Not Agree on the True Model

STEPHEN J. TURNOVSKY, TAMER BASAR, AND VASCO D'OREY

Dynamic Strategic Monetary Policies and Coordination  
in Interdependent Economies.

ALAN C. STOCKMAN AND ALEJANDRO HERNÁNDEZ D.

Exchange Controls, Capital Controls, and  
International Financial Markets

LARS E. O. SVENSSON

Trade in Risky Assets

HARVEY E. LAPAN The Optimal Tariff, Production Lags, and Time Consistency

JEREMY GREENWOOD, ZVI HERCOWITZ, AND GREGORY W. HUFFMAN

Investment, Capacity Utilization, and the Real  
Business Cycle

ANGUS DEATON. Quality, Quantity, and Spatial Variation of Price

ALLEN DRAZEN AND ZVI ECKSTEIN

On the Organization of Rural Markets and the Process  
of Economic Development

LEWIS EVANS AND STEVEN GARBER

Public-Utility Regulators Are Only Human: A Positive  
Theory of Rational Constraints.

GRAHAM LOOMES When Actions Speak Louder Than Prospects

GUR HUBERMAN AND CHARLES KAHN

Limited Contract Enforcement and Strategic Renegotiation

VINCENT CRAWFORD

Long-Term Relations Governed by Short-Term Contracts

YORAM KROLL, HAIM LEVY, AND AMNON RAPOPORT

Experimental Tests of the Separation Theorem and the  
Capital Asset Pricing Model

SHORTER PAPERS: B. T. Diba and H. I. Grossman; I. L. Collier, Jr. and D. H. Papell; E. B. Field;  
F. R. Lichtenberg; K. A. Small and C. Winston; W. R. Johnson; N. Kulatilaka and S. G. Marks;  
S. Awerbuch.

JUNE 1988

# THE AMERICAN ECONOMIC ASSOCIATION

P14422

●Printed at Banta Company, Menasha, Wisconsin.

●No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, subscriptions, and changes of address, should be sent to the American Economic Association, 1313 21st Avenue South, Suite 809, Nashville, TN 37212-2786. Please remit membership payment with the application included elsewhere in this journal. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

*THE AMERICAN ECONOMIC REVIEW* (ISSN 0002-8282), June 1988, Vol. 78, No. 3, is published five times a year (March, May, June, September, December) by the American Economic Association, 1313 21st Avenue South, Suite 809, Nashville, TN 37212-2786. Annual subscription fees: Institutional subscriber-\$125.00, Individual subscriber-\$72.00, Regular members-\$38.50, \$46.20, or \$53.90 depending on income. A subscription also includes the *Journal of Economic Literature* and the *Journal of Economic Perspectives*. In countries other than the U.S.A., add \$16.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *American Economic Review*, 1313 21st Avenue South, Suite 809, Nashville, TN 37212-2786.

Founded in 1885

## Officers

### *President*

ROBERT EISNER  
Northwestern University

### *President-elect*

JOSEPH A. PECHMAN  
The Brookings Institution

### *Vice-Presidents*

MARTIN S. FELDSTEIN  
National Bureau of Economic Research  
and Harvard University  
F. M. SCHERER  
Swarthmore College

### *Secretary-Treasurer*

C. ELTON HINSHAW  
Vanderbilt University

### *Editor of The American Economic Review*

ORLEY C. ASHENFELTER  
Princeton University

### *Editor of The Journal of Economic Literature*

JOHN PENCABEL  
Stanford University

### *Editor of The Journal of Economic Perspectives*

JOSEPH E. STIGLITZ  
Princeton University

## Executive Committee

### *Elected Members of the Executive Committee*

SHERWIN ROSEN  
University of Chicago  
THOMAS J. SARGENT  
University of Minnesota  
ROBERT J. BARRO  
Harvard University  
JUDITH A. THORNTON  
University of Washington  
GEORGE A. AKERLOF  
University of California-Berkeley  
ISABEL V. SAWHILL  
Urban Institute

### *EX OFFICIO Members*

ALICE M. RIVLIN  
The Brookings Institution  
GARY S. BECKER  
University of Chicago





Arthur S. Feldberger

391  
1001

## ARTHUR S. GOLDBERGER

DISTINGUISHED FELLOW

1988

Arthur S. Goldberger has been a true and exemplary scholar in quantitative methods, both for economics and its sister social sciences. From the very beginning of his professional career, he has been deeply concerned with the accuracy of economic measurements and the standard techniques of econometric model building. His early applications and analyses of empirical models brought understanding to a whole generation of econometricians.

As he matured professionally he mastered econometric theory to a point where he not only made creative discoveries but also prepared lucid explanations for students and practitioners. He mastered the linear multivariate statistical model. He was then led into frontier territory where econometricians had feared to tread, and through both insight and perseverance showed how models with errors of measurement and unobservables could be estimated. This took him away from the strict field of econometrics into other social sciences where the factor analysis and latent variable models were being fruitfully used. His methodological and substantive findings, also in the area of behavior genetics, stand as everlasting contributions of one who has successfully bridged the many gaps that separate the social disciplines. The American Economic Association honors Arthur S. Goldberger as the quintessential interdisciplinary social scientist.

# THE AMERICAN ECONOMIC REVIEW

Editor

ORLEY ASHENFELTER

Co-Editors

ROBERT H. HAVEMAN

JOHN B. TAYLOR

HAL R. VARIAN

Production Editor

CLAIRE H. COMISKEY

Board of Editors

GEORGE A. AKERLOF

JAMES E. ANDERSON

JO ANNA GRAY

GEORGE E. JOHNSON

KENNETH L. JUDD

JOHN F. KENNAN

MAURICE OBSTFELD

ROBERT H. PORTER

JOHN G. RILEY

RICHARD ROLL

ALVIN E. ROTH

DAVID SAPPINGTON

KENNETH J. SINGLETON

ROBERT S. SMITH

BARBARA J. SPENCER

LESLIE YOUNG

•Submit manuscripts (4 copies), 50 pages maximum, double-spaced, to:

Orley Ashenfelter, Editor, *AER*; 209 Nassau Street, Princeton, NJ 08542-4607.

•Submission fee: \$50 for members; \$100 for nonmembers. Please pay with a check or money order payable in United States Dollars. Canadian and Foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Style guides will be provided upon request.

•Copyright © American Economic Association 1988. All rights reserved.

June 1988

VOLUME 78, NUMBER 3

Articles

- Growth Theory and After *Robert M. Solow* 307
- International Macroeconomic Policy Coordination When Policymakers Do Not Agree on the True Model *Jeffrey A. Frankel and Katharine Rockett* 318
- Dynamic Strategic Monetary Policies and Coordination in Interdependent Economies *Stephen J. Turnovsky, Tamer Baser, and Vasco d'Orey* 341
- Exchange Controls, Capital Controls, and International Financial Markets *Alan C. Stockman and Alejandro Hernández D.* 362
- Trade in Risky Assets *Lars E. O. Svensson* 375
- The Optimal Tariff, Production Lags, and Time Consistency *Harvey E. Lapan* 395
- Investment, Capacity Utilization, and the Real Business Cycle *Jeremy Greenwood, Zvi Hercowitz, and Gregory W. Huffman* 402
- Quality, Quantity, and Spatial Variation of Price *Angus Deaton* 418
- On the Organization of Rural Markets and the Process of Economic Development *Allan Drazen and Zvi Eckstein* 431
- Public-Utility Regulators Are Only Human: A Positive Theory of Rational Constraints *Lewis Evans and Steven Garber* 444
- When Actions Speak Louder Than Prospects *Graham Loomes* 463
- Limited Contract Enforcement and Strategic Renegotiation *Gur Huberman and Charles Kahn* 471
- Long-Term Relationships Governed by Short-Term Contracts *Vincent Crawford* 485
- Experimental Tests of the Separation Theorem and the Capital Asset Pricing Model *Yoram Kroll, Haim Levy, and Amnon Rapoport* 500

## Shorter Papers

Explosive Rational Bubbles in Stock Prices?	<i>Behzad T. Diba and Herschel I. Grossman</i>	520
About Two Marks: Refugees and the Exchange Rate Before the Berlin Wall	<i>Irwin L. Collier, Jr. and David H. Papell</i>	531
The Relative Efficiency of Slavery Revisited: A Translog Production Function Approach	<i>Elizabeth B. Field</i>	543
The Private <i>R&amp;D</i> Investment Response to Federal Design and Technical Competitions	<i>Frank R. Lichtenberg</i>	550
Optimal Highway Durability	<i>Kenneth A. Small and Clifford Winston</i>	560
Income Redistribution in a Federal System	<i>William R. Johnson</i>	570
The Strategic Value of Flexibility: Reducing the Ability to Compromise	<i>Nalin Kulatilaka and Stephen Gary Marks</i>	574
Accounting Rates of Return: Comment	<i>Shimon Awerbuch</i>	581
<b>Auditors' Report</b>		<b>588</b>

# Growth Theory and After<sup>†</sup>

By ROBERT M. SOLOW\*

I have been told that everybody has dreams, but that some people habitually forget them even before they wake up. That seems to be what happens to me. So I do not know if I have ever dreamed about giving this lecture. I know I have been in this room before, but that was in real life, and I was awake. If I have given this lecture in my dreams, there is no doubt that the topic was the theory of economic growth. I am told that the subject of the lecture should be "on or associated with the work for which the Prize was awarded." That is pretty unambiguous. But I would not even wish to use the leeway offered by the phrase "associated with." Growth theory is exactly what I want to talk about: for itself, for its achievements, for the gaps that remain to be filled, and also as a vehicle for some thoughts about the nature of theoretical research in macroeconomics, and empirical research as well.

Growth theory did not begin with my articles of 1956 and 1957, and it certainly did not end there. Maybe it began with *The Wealth of Nations*; and probably even Adam Smith had predecessors. More to the point, in the 1950s I was following a trail that had been marked out by Roy Harrod and by Evsey Domar, and also by Arthur Lewis in a slightly different context. Actually I was trying to track down and relieve a certain discomfort that I felt with their work. I shall try to explain what I mean in a few words.

Harrod and Domar seemed to be answering a straightforward question: When is an economy capable of steady growth at a con-

stant rate? They arrived by noticeably different routes, at a classically simple answer: the national saving rate (the fraction of income saved) has to be equal to the product of the capital-output ratio and the rate of growth of the (effective) labor force. Then and only then could the economy keep its stock of plant and equipment in balance with its supply of labor, so that steady growth could go on without the appearance of labor shortage on one side or labor surplus and growing unemployment on the other side. They were right about that general conclusion.

Discomfort arose because they worked this out on the assumption that all three of the key ingredients—the saving rate, the rate of growth of the labor force, and the capital-output ratio—were given constants, facts of nature. The saving rate was a fact about preferences; the growth rate of labor supply was a demographic-sociological fact; the capital-output ratio was a technological fact.

All of them were understood to be capable of changing from time to time, but sporadically and more or less independently. In that case, however, the possibility of steady growth would be a miraculous stroke of luck. Most economies, most of the time, would have no equilibrium growth path. The history of capitalist economies should be an alternation of long periods of worsening unemployment and long periods of worsening labor shortage.

The theory actually suggested something even more dramatic. Harrod's writings, especially, were full of incompletely worked out claims that steady growth was in any case a very unstable sort of equilibrium: any little departure from it would be magnified indefinitely by a process that seemed to depend mainly on vague generalizations about entrepreneurial behavior. You may remember that John Hicks's *Trade Cycle* book, which was based on Harrod's growth model, needed to invoke a full employment ceiling to generate downturns and a zero-gross-investment

<sup>†</sup>This is the lecture Robert Solow delivered in Stockholm, Sweden, December 8, 1987, when he received the Nobel Prize in Economic Science. The article is copyright © The Nobel Foundation 1987, and published here with the permission of The Nobel Foundation.

\*Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139.

floor to generate upturns. Otherwise the model economy would have run away.

Keep in mind that Harrod's first *Essay* was published in 1939 and Domar's first article in 1946. Growth theory, like much else in macroeconomics, was a product of the depression of the 1930s and of the war that finally ended it. So was I. Nevertheless it seemed to me that the story told by these models felt wrong. An expedition from Mars arriving on Earth, having read this literature, would have expected to find only the wreckage of a capitalism that had shaken itself to pieces long ago. Economic history was indeed a record of fluctuations as well as of growth, but most business cycles seemed to be self-limiting. Sustained, though disturbed, growth was not a rarity.

There was another implication of the Harrod-Domar model that seemed unsound. If the condition for steady growth in a labor-surplus economy is that the savings rate equal the product of the growth rate of employment and a technologically determined capital-output ratio, then a recipe for doubling the rate of growth was simply to double the savings rate, perhaps through the public budget. Well, not *simply*: we all knew then—as I am not sure we all know now—that doubling the *ex ante* saving rate would not double the *ex post* saving rate unless something were taking care of the *ex ante* investment rate at the same time. (I hope these strange Latin phrases are still understood in Stockholm in 1987!) In underdeveloped countries, however, where the appetite for new capital is likely to be pretty strong, the recipe looked usable. I believe I remember that writings on economic development often asserted that the key to a transition from slow growth to fast growth was a sustained rise in the savings rate. The recipe sounded implausible to me. I can no longer remember exactly why, but it did.

That was the spirit in which I began tinkering with the theory of economic growth, trying to improve on the Harrod-Domar model. I cannot tell you why I thought first about replacing the constant capital-output (and labor-output) ratio by a richer and more realistic representation of the technology. I know that even as a student I was drawn to

the theory of production rather than to the formally almost identical theory of consumer choice. It seemed more down to earth. I know that it occurred to me very early, as a natural-born macroeconomist, that even if technology itself is not so very flexible for each single good at a given time, aggregate factor intensity must be more variable because the economy can choose to focus on capital-intensive or labor-intensive or land-intensive goods. Anyway, I found something interesting right away.

It would sound silly for me to explain in any detail to this audience what I found. Nearly everyone who spends any time in this room already knows. The "neoclassical model of economic growth" started a small industry. It stimulated hundreds of theoretical and empirical articles by other economists. It very quickly found its way into textbooks and into the fund of common knowledge of the profession. Indeed that is what allows me to think that I am a respectable person to be giving this lecture today. Nevertheless I must summarize the outcome in a couple of sentences, so that I can move on to the more interesting questions about what is still unknown or uncertain and remains to be found out.

Just allowing for a reasonable degree of technological flexibility accomplished two things. In the first place, the mere existence of a feasible path of steady growth turned out not to be a singular event. A range of steady states is possible, and the range may even be quite wide if the range of aggregative factor intensities is wide. There are other ways in which an economy can adapt to the Harrod-Domar condition, but it still seems to me that variation in capital intensity is probably the most important.

Second, it turned out to be an implication of diminishing returns that the equilibrium rate of growth is not only not proportional to the saving (investment) rate, but is independent of the saving (investment) rate. A developing economy that succeeds in permanently increasing its saving (investment) rate will have a higher level of output than if it had not done so, and must therefore grow faster for a while. But it will not achieve a permanently higher rate of growth of output.



More precisely, the permanent rate of growth of output per unit of labor input is independent of the saving (investment) rate and depends entirely on the rate of technological progress in the broadest sense.

There was a third result that seemed useful and certainly helped to make the model appealing to economists. Earlier growth theory was mechanical or physical, not in any bad sense but in the sense that it was almost entirely a description of flows and stocks of goods. In the neoclassical model it was quite natural and practical to describe *equilibrium* paths and to work out the price and interest rate dynamics that would support an equilibrium path. It did not occur to me at the time that in doing this I was bringing good news and bad news. The good news was that economists instinctively like to think that way, and the connection would help to get my professional colleagues interested in growth theory. Moreover, it is a good (that is, fruitful) instinct, whether one is dealing with a capitalist or a socialist economy. The bad news is that the connection is a bit too pretty and too interesting and unleashes a standing temptation to sound like Dr. Pangloss, a very clever Dr. Pangloss. I think that tendency has won out in recent years, as I shall try to explain later on, though it may be too late for me to pretend to be Candide.

When I look back now at the articles I wrote in the 1950s and 1960s on this general subject, I am struck and even a little surprised at how much effort went into broadening the technological framework of growth theory. I wanted to make sure that the model could accommodate the likelihood that new technology can only be introduced with the use of newly designed and produced capital equipment, that factor proportions might be variable only at the instant of gross investment and not after capital equipment had taken some particular form, and that enough flexibility could be achieved with discrete activities, even with only one activity so long as the length of life of capital goods could be chosen economically. And in every case I wanted to show that the appropriate commodity-price-factor-price relations could be worked out and made intelligible in terms of the inherited instincts of economists. (In my

case I had inherited them mainly from Knut Wicksell and Paul Samuelson.)

There were reasons for this special orientation, reasons that seemed pretty compelling at the time. In the first place, it was the introduction of some technological flexibility that had opened up growth theory to a wider variety of real-world facts and to a closer connection with general economic theory. It seemed important to make sure that these gains were not tied too closely to an indefensibly simple version of factor substitution. Second, I had already begun to do some empirical work making use of an aggregate production function with apparently meaningful and clearly surprising results. I was very skeptical about this device myself, and I knew that others would have doubts of their own. It seemed like a good idea to make sure that the method was capable, at least in principle, of dealing with the first few doses of realism. And, third, I was already trapped in the famous "Cambridge controversy." I use the word "trapped" because that whole episode now seems to me to have been a waste of time, a playing-out of ideological games in the language of analytical economics. At the time I thought—and the literature gave some reason to think—that part of the argument was about marginalism, about smooth marginalism. So I wanted to be able to show that the conclusions of the theory and of its empirical implementation were not bound to that very special formulation. I guess it was worth doing, but it certainly did not pacify anyone.

There was one bad by-product of this focus on the description of technology. I think I paid too little attention to the problems of effective demand. To put it differently; a theory of equilibrium growth badly needed—and still needs—a theory of deviations from the equilibrium growth path. I can honestly say that I realized the need at the time. There is a brief section at the end of my 1956 article that deals in a perfunctory way with the implications of real-wage rigidity and with the possibility of a liquidity trap. That was just a lick and a promise. There was also a paragraph that I am prouder of: it made the point that growth theory provides a framework within which one can

seriously discuss macroeconomic policies that not only achieve and maintain full employment but also make a deliberate choice between current consumption and current investment, and therefore between current consumption and future consumption. Only a few years later I had the memorable experience in the Kennedy-Heller Council of Economic Advisers of seeing those ideas written into the 1962 *Economic Report* (which is about to be republished by the MIT Press). The history of the past seven years in the United States suggests that the lesson has not yet been learned in Washington.

The problem of combining long-run and short-run macroeconomics has still not been solved. I will come back to it later on. This is the place for me to confess to (and explain away) a certain youthful confusion. In the early discussions of Harrod-Domar growth theory there was much talk about the intrinsic instability of equilibrium growth. "Instability" could and did mean two different things, and the meanings were not always clearly distinguished. It could mean that well-behaved equilibrium paths are surrounded by badly behaved equilibrium paths, so that a small sideward step could lead to eventual disaster. Or it could mean that instability applies to disequilibrium behavior, so that an economy that once strays from equilibrium growth would not automatically find its way back to *any* equilibrium growth path.

The original Harrod-Domar model seemed to be subject to both these difficulties. I think I showed that extension of the model took the sting out of the first sort of instability. The second sort, however, really does involve the integration of short-run and long-run macroeconomics, of growth theory and business-cycle theory. Harrod and many contemporary commentators went at this problem by making very special (and unconvincing) assumptions about investment behavior. I may not have been as clear then as I am now about the distinction between the two notions of instability. Today I would put the unsolved problem as follows. One of the achievements of growth theory was to relate equilibrium growth to asset pricing

under tranquil conditions. The hard part of disequilibrium growth is that we do not have—and it may be impossible to have—a really good theory of asset valuation under turbulent conditions. (1987 is an excellent year in which to make that observation!)

One important tendency in contemporary macroeconomic theory evades this problem in an elegant but (to me) ultimately implausible way. The idea is to imagine that the economy is populated by a single immortal consumer, or a number of identical immortal consumers. The immortality itself is not a problem—each consumer could be replaced by a dynasty, each member of which treats her successors as extensions of herself. But no shortsightedness can be allowed. This consumer does not obey any simple short-run saving function, or even a stylized Modigliani life-cycle rule of thumb. Instead she, or the dynasty, is supposed to solve an infinite-time utility-maximization problem. That strikes me as farfetched, but not so awful that one would not want to know where the assumption leads.

The next step is harder to swallow in conjunction with the first. For this consumer every firm is just a transparent instrumentality, an intermediary, a device for carrying out intertemporal optimization subject only to technological constraints and initial endowments. Thus any kind of market failure is ruled out from the beginning, by assumption. There are no strategic complementarities, no coordination failures, no Prisoners Dilemmas.

The end result is a construction in which the whole economy is assumed to be solving a Ramsey optimal-growth problem through time, disturbed only by stationary stochastic shocks to tastes and technology. To these the economy adapts optimally. Inseparable from this habit of thought is the automatic presumption that observed paths are equilibrium paths. So we are asked to regard the construction I have just described as a model of the actual capitalist world. What we used to call business cycles—or at least booms and recessions—are now to be interpreted as optimal blips in optimal paths in response to random fluctuations in productivity and the desire for leisure.



I find none of this convincing. The markets for goods and labor look to me like imperfect pieces of social machinery with important institutional peculiarities. They do not seem to behave at all like transparent and frictionless mechanisms for converting the consumption and leisure desires of households into production and employment decisions. I cannot imagine shocks to taste and technology large enough on a quarterly or annual time scale to be responsible for the ups and downs of the business cycle. But now I have to report something disconcerting. I can refer you to an able, civilized and completely serious example of this approach and suggest that you will find it very hard to refute. You can find nontrivial objections to important steps in the argument, but that would be true of any powerful macroeconomic model.

There is a dilemma here. When I say that E. Prescott's story is hard to refute, it does not follow that his case can be proved. Quite the contrary: there are other models, inconsistent with his, that are just as hard to refute, maybe harder. The conclusion must be that historical time-series do not provide a critical experiment. This is where a chemist would move into the laboratory, to design and conduct just such an experiment. That option is not available to economists. My tentative resolution of the dilemma is that we have no choice but to take seriously our own direct observations of the way economic institutions work. There will, of course, be arguments about the *modus operandi* of different institutions, but there is no reason why they should not be intelligible, orderly, fact-bound arguments. This sort of methodological opportunism can be uncomfortable and unsettling; but at least it should be able to protect us from foolishness.

Since what I have just said goes against the spirit of the times, I would like to be very explicit. No one could be against time-series econometrics. When we need estimates of parameters, for prediction or policy analysis, there is no good alternative to the specification and estimation of a model. To leave it at that, however, to believe as many American economists do that empirical economics begins and ends with time-series

analysis, is to ignore a lot of valuable information that cannot be put into so convenient a form. I include the sort of information that is encapsulated in the qualitative inferences made by expert observers, as well as direct knowledge of the functioning of economic institutions. Skepticism is always in order, of course. Insiders are sometimes the slaves of silly ideas. But we are not so well off for evidence that we can afford to ignore everything but time-series of prices and quantities.

After this methodological digression, I should remind you of the direction of my main argument. Growth theory was invented to provide a systematic way to talk about and to compare equilibrium paths for the economy. In that task it succeeded reasonably well. In doing so, however, it failed to come to grips adequately with an equally important and interesting problem: the right way to deal with deviations from equilibrium growth. One possible solution strikes me as wrongheaded: that is to deny the existence of an analytical problem by claiming that "economic fluctuations" are not deviations from equilibrium growth at all, but examples of equilibrium growth. My impression is that belief in this story is more or less confined to North America and perhaps the Federal Republic of Germany. Maybe the experiences of other European economies do not lend themselves to this interpretation at all. What alternatives are there?

It will not do simply to superimpose your favorite model of the business cycle on an equilibrium growth path. That might do for very small deviations, more in the nature of minor slightly autocorrelated "errors." But if one looks at substantial more-than-quarterly departures from equilibrium growth, as suggested for instance by the history of the large European economies since 1979, it is impossible to believe that the equilibrium growth path itself is unaffected by the short- to medium-run experience. In particular the amount and direction of capital formation is bound to be affected by the business cycle, whether through gross investment in new equipment or through the accelerated scrapping of old equipment. I am also inclined to believe that the segmentation of the labor

market by occupation, industry, and region, with varying amounts of unemployment from one segment to another, will also react back on the equilibrium path. So a simultaneous analysis of trend and fluctuations really does involve an integration of long run and short run of equilibrium and disequilibrium.

The simplest strategy is a familiar one from other contexts. In a completely aggregated growth model the relevant prices are the real wage and real rate of interest. Suppose they are both rigid, or merely adjust very slowly to excess supplies in the markets for labor and goods. (The more usual assumption is that only the wage is sticky; but in Knut Wicksell's own native habitat we should allow for a divergence between the "natural" and "market" rates of interest.) Then the economy may be away from any full-equilibrium path for a long time. During that time its evolution will be governed by a short-run dynamics much like everyday business-cycle theory.

The most interesting case to consider is one where real wage and rate of interest are stuck at levels that lead to excess supply of labor and goods (saving greater than investment *ex ante*). This is the sort of configuration we have come to call "Keynesian." The big difference is that net investment may be positive or negative; industrial capacity may be rising or falling. The economy may eventually return to an equilibrium path, perhaps because "prices are flexible in the long run" as we keep telling ourselves. If and when it does, it will not return to the continuation of the equilibrium path it was on before it slipped off. The new equilibrium path will depend on the amount of capital accumulation that has taken place during the period of disequilibrium, and probably also on the amount of unemployment, especially long-term unemployment, that has been experienced. Even the level of technology may be different, if technological change is endogenous rather than arbitrary.

This is the sort of amendment that I mentioned in 1956, but did not pursue very far. There is now an excellent exploratory sketch by Edmond Malinvaud using this fix-price approach to growth theory. As you would expect, an important role is played by the investment function. When I referred earlier

on to the difficult problem of asset valuation away from an equilibrium path, this is what I meant. We are reduced to some more or less plausible formulation guided by more or less robust econometrics results and by whatever we think we know about investment decision making in real firms. Malinvaud emphasizes "profitability" as a determinant of investment, but he also emphasizes that the precise meaning of profitability is unclear whenever the future is unclear.

The main result of Malinvaud's analysis is a clarification of the condition under which a "Keynesian" steady state is possible, and when it is locally stable, that is, when it will be approached by an economy disturbed from a nearby equilibrium path. The unstable case is just as interesting, because it suggests the possibility of small causes having big results. All these stability arguments have to be tentative because the interest rate and real wage are assumed to be fixed while quantities move. That is not an adequate reason to dismiss the results in a purist spirit; but obviously the research program is not complete.

A sketch by Malinvaud is as good as a book by someone else. My own inclination—it is just an inclination—is to try a slightly different slant. Thinking about the ambiguity of the concept of profitability and its relation to investment reminds one that many firms react to changed circumstances precisely by changing their prices. The obvious alternative to a model with sticky prices is a model with imperfectly competitive price-setting firms. Then, of course, one can no longer speak in any simple way of excess supply of goods. But we can find something just as interesting; the possibility of many coexisting equilibrium paths, some of which are unambiguously better than others. (Usually the better ones have higher output and employment than the worse ones, so something like recession makes an appearance anyway.) The interaction of growth and business cycle can then take a slightly different form: alternation of good and bad equilibria is not just a simple averaging.

This sort of model is now pretty familiar in a static context, where it can make good working sense of the notion of "effective demand." Firms will naturally condition

their actions on beliefs about economic aggregates. Frank Hahn and I are working on extending it to a model of overlapping generations, so that it would be easy to convert any stationary equilibrium state into a growing steady state. Preliminary indications are that the thing can be done. There is a hope, therefore, that either the fix-price approach or the imperfect-competition approach can allow us to talk sensibly about macroeconomic policy in a growth context.

In my 1956 paper there was already a brief indication of the way neutral technological progress could be incorporated into a model of equilibrium growth. It was a necessary addition because otherwise the only steady states of the model would have constant income per person and that could hardly be a valid picture of industrial capitalism. Technological progress, very broadly defined to include improvements in the human factor, was necessary to allow long-run growth in real wages and the standard of living. Since an aggregate production function was already part of the model, it was natural to think of estimating it from long-run time-series for a real economy. That plus a few standard parameters—like saving rate and population growth—would make the model operational.

Estimating an aggregate production function was hardly a new idea, but I did have a new wrinkle in mind: to use observed factor prices as indicators of current marginal productivities, so that each observation would give me not only an approximate point on the production function but also an approximate indication of its slopes. I am pretty sure that this idea was suggested to me by equilibrium growth theory. I want to emphasize that I did not then have any notion that I was doing something intensely controversial.

The first few paragraphs of my 1957 article are thoroughly ambivalent, not about the method but about the use of aggregate data on inputs and output. After expressing my doubts I went ahead in a pragmatic spirit. One cannot do macroeconomics without aggregative relationships; and at least for the moment there is no substitute for macroeconomics. The only way I can account for the intensity of controversy over this point is to

ascribe it to the belief that there is something intrinsically ideological about the notion that profits on "capital" represents the return to a factor of production as imputed by the market. John Bates Clark may have thought, a century ago, that distribution according to marginal products was "just" but no modern economist, no modern "bourgeois" economist, would accept that reasoning.

Anyway, the main result of that 1957 exercise was startling. Gross output per hour of work in the U.S. economy doubled between 1909 and 1949; and some seven-eighths of that increase could be attributed to "technical change in the broadest sense" and only the remaining eighth could be attributed to conventional increase in capital intensity. Actually Solomon Fabricant at the National Bureau of Economic Research had come up with a similar breakdown for a slightly earlier period, using methods with less in the way of analytic foundation. I think I had expected to find a larger role for straightforward capital formation than I actually found; I will come back to that point soon.

The broad conclusion has held up surprisingly well in the thirty years since then during which time "growth accounting" has been refined quite a lot, especially by Edward Denison (1985). The main refinement has been to unpack "technical progress in the broadest sense" into a number of constituents of which various human-capital variables and "technological change in the narrow sense" are the most important. To give you an idea of the current state of play I shall quote Denison's most recent estimates for the United States.

Taking the period from 1929 to 1982 and smoothing away the business cycle, he finds that real nonresidential business output increased at an average rate of 3.1 percent a year. The problem now is to parcel this out among a number of basic determinants of growth. Denison (1985) estimates that a quarter of it can be attributed to increased labor input of constant educational level. Another 16 percent (i.e., about one-half percent a year) is credited to the increased educational qualifications of the average worker. The growth of "capital" accounts

for 12 percent of the growth of output; this is coincidentally almost exactly what I found for 1909–49 using my original method, of which Denison's (1985) is in some ways a practical refinement. Then Denison imputes 11 percent of total growth to "improved allocation of resources" (by which he means such things as the movement of labor from low-productivity agriculture to higher productivity industry). Another 11 percent goes to "economies of scale" (but this must be a very insecure imputation). Finally, 34 percent of recorded growth is credited to "the growth of knowledge" or technological progress in the narrow sense. If you add up these percentages, you will see that Denison has accounted for 109 percent of measured growth. Miscellaneous factors must then have reduced the growth of output by nine percent of 3.1 percent, or just under 0.3 percent a year. (These negative factors could include such things as investment in environmental improvement, which uses resources but does not appear in measured output, though it may of course be very valuable.)

This detailed accounting is an improvement on my first attempt, but it leads to roughly the same conclusion. Remember that I distinguished only three factors: straight labor, straight capital, and residual "technical change." Denison (1985) decomposes the residual into five components, but the flavor is very similar.

The similarity is brought out more strongly if one looks at Denison's results on a "per person employed" basis. Real output per person employed grew by 1.7 percent per year between 1929 and 1982. Labor input per person employed accounted for –23 percent of this. That sounds strange; but means mostly that hours worked per year per person employed fell during this period, so that the average employed person provided less straight labor time. I will not go over the full imputation. All I want to point out is that education per worker accounts for 30 percent of the increase in output per worker and the advance of knowledge accounts for 64 percent in Denison's figures. Thus technology remains the dominant engine of growth, with human capital investment in second place. One does not have to

believe in the accuracy of these numbers; the message they transmit is pretty clear anyway.

That is meant as a serious remark. If I may revert to methodological propaganda again, I would like to remind my colleagues and their readers that every piece of empirical economics rests on a substructure of background assumptions that are probably not quite true. For instance, these total-factor-productivity calculations require not only that market prices can serve as a rough-and-ready approximation of marginal products, but that aggregation does not hopelessly distort these relationships. Under those circumstances, robustness should be the supreme econometric virtue; and over-interpretation is the endemic econometric vice. So I would be happy if you were to accept that the results I have been quoting point to a qualitative truth and give perhaps some guide to orders of magnitude. To ask for much more than that is to ask for trouble. I would also like to quote the profound warning issued by the leading student of the statistics of baseball—it hangs in my office—"No amount of (apparent) statistical evidence will make a statement invulnerable to common sense."

The mention of common sense brings to mind another aspect of this story, still unsettled in the literature. In the beginning, I was quite surprised at the relatively minor part the model ascribed to capital formation. Even when this was confirmed by Denison and others, the result seemed contrary to common sense. The fact that the steady-state rate of growth is independent of the investment quota was easy to understand; it only required thinking through the theory. It was harder to feel comfortable with the conclusion that even in the shorter run increased investment would do very little for transitory growth. The transition to a higher equilibrium growth path seemed to offer very little leverage for policy aimed at promoting investment.

The formal model omitted one mechanism whose absence would clearly bias the predictions against investment. That is what I called "embodiment," the fact that much technological progress, maybe most of it, could find

its way into actual production only with the use of new and different capital equipment. Therefore the effectiveness of innovation in increasing output would be paced by the rate of gross investment. A policy to increase investment would thus lead not only to higher capital intensity, which might not matter much, but also to a faster transfer of new technology into actual production, which would. Steady-state growth would not be affected, but intermediate-run transitions would, and those should be observable.

That idea seemed to correspond to common sense, and it still does. By 1958 I was able to produce a model that allowed for the embodiment effect. A certain amount of simplicity was lost, because the stock of capital could no longer be regarded as a homogeneous lump. One had to keep track of its age structure; but that was precisely the point. Anyhow the model was workable even if it was not neat. If common sense was right, the embodiment model should have fit the facts significantly better than the earlier one. But it did not. Denison (1985), whose judgment I respect, came to the conclusion that there was no explanatory value in the embodiment idea. I do not know if that finding should be described as a paradox, but it was at least a puzzle.

In the course of preparing this lecture, I came across a recent working paper by Edward N. Wolff (1987), of New York University, which offers a longer-run perspective on this matter. Wolff compiled data for seven large countries (Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States) covering the whole century from 1880 to 1979. He also paid special attention to the postwar period 1950–79. These particular countries were selected for data availability only, so they cannot be considered a representative sample. Wolff's result is therefore only suggestive, but it is an interesting suggestion.

For each of the countries he calculates the average growth rate of total factor productivity (i.e., what I have called the rate of technical progress in the broad sense) and also various measures of the speed of investment. (For instance, he looks at the growth rate of the capital stock, the growth rate of

the capital-labor ratio, and the average investment quota itself.) Then, looking across countries, he finds a very strong positive correlation between the rate of technical progress and the speed of investment. His interpretation is that this provides strong confirmation of the embodiment hypothesis: if we suppose that all these countries had access to roughly the same pool of technological innovations, then it appears that the ones that invested fastest were best able to take advantage of the available knowledge. That is certainly one reasonable interpretation and it is one I like. Keep in mind that, by using total factor productivity, Wolff has already "given" to investment its traditional function of increasing productivity by increasing capital intensity, so the remaining correlation is between investment and the *shift* of the aggregate production function.

To be faithful to my own methodological precepts, however, I should remind you that other interpretations are also possible. For example, it could be the case that some countries are better able to exploit the common pool of technological progress than others, for reasons that have nothing to do with the rate of capital formation; but in exactly those technologically progressive countries investment is most profitable, so naturally the rate of investment is higher. Or else rapid technical progress *and* high investment could both be the result of some third factor, like the presence of conditions that encourage entrepreneurial activity. High investment and fast technical progress will then go together.

I cannot argue strongly one way or the other. But at least the way remains open for a reasonable person to believe that the stimulation of investment will favor faster intermediate-run growth through its effect on the transfer of technology from laboratory to factory.

Before I finish, perhaps I should point out that it is possible to combine most of the building blocks I have been discussing in a small but fairly complete econometric model. If that were not possible, I would find the ideas less interesting. It has in fact been done. One example is the "annual growth model of the U.S. economy" due to Bert

Hickman and Robert Coen (1976). This is a model whose production side is completely aggregated and is, in fact, just exactly the sort of thing I have been talking about. (The demand side is disaggregated, but that is not important now.) The full-equilibrium paths of the Hickman-Coen (1976) model are exactly those made familiar by growth theory, a little more general because the determination of saving and the evolution of the labor force are looked after in more detail.

That part is quite straightforward. In some recent exercises, however, Hickman (1987) has started a serious study of deviations from equilibrium growth in exactly the spirit recommended by Malinvaud (1983) and by me. He allows for real wage rigidity, and then models the producing sector as a price-setting monopolistic competitor. Now investment does not have to be equal to full-employment saving, except in full equilibrium. Periods of boom and stagnation can appear, and do appear, to almost no one's surprise. There can be "Keynesian" and "classical" unemployment. Indeed there can be both at the same time: The real wage might be too high to allow full employment with existing capital stock, while at the same time aggregate demand is inadequate to take off the market what firms would wish to produce. Changes in the real wage could have demand-side and supply-side effects.

All this sounds very good, sounds just like the macroeconomics that pragmatic Americans and Swedes have practiced all along. I cannot vouch for the Hickman numbers, but they are at least sensible. They show, by the way, that high real-wage induced unemployment was negligible in the United States between 1959 and 1978, and was then again dwarfed by low demand-induced unemployment in 1981 and 1982. I do not know what their story is for the years after 1982, but the fact that I would like to know speaks well for the model.

In this brief review of the goals and achievements of growth theory I have referred as much to the work of others as to my own. That is more than mere modesty: the choice reflects my belief that any successful line of economic analysis is almost

certain to be a group product. We attach names to ideas for good and bad reasons, but useful ideas are usually worked out and critically refined by a research community. I have some faith that the ideas of "neoclassical" growth theory are viable just because they have attracted a research community, even a rather diverse community: Robert Lucas and Prescott build on the basic model, and so do Malinvaud and "sunspot" theorists like Karl Shell and others.

When I read Robert Frost's lines from "The Black Cottage":

Most of the change we think we see in  
life is due to truths being in and out of  
favor.

it occurred to me at once that they sound altogether too much like economics. Some of that feeling is inevitable, and not necessarily to be regretted. The permanent substructure of applicable economics cannot be too very large because social institutions and social norms evolve, and the characteristics of economic behavior will surely evolve with them. I believe also that part of the changeability of economic ideas on a shorter time scale is our own doing. It comes from trying too hard, pushing too far, asking evermore refined questions of limited data, over-fitting our models and over-interpreting the results. This, too, is probably inevitable and not especially to be regretted. You never know if you have gone as far as you can until you try to go further.

Naturally I hope that growth theory can serve in both ways: as a background on which to hang multisector models that probably try to do more than can be done, and as a framework for simple, strong, loosely quantitative propositions about cause and effect in macroeconomics. For both roles, it appears to me, the fundamental intellectual need is for a common understanding of medium-run departures from equilibrium growth. That is the stuff of everyday macroeconomics. It has been going on in English-speaking countries since Keynes and in Sweden since Lindahl and the Stockholm School. It is going on in both places today.

## REFERENCES

- Denison, E., *Trends in American Economic Growth, 1929-1982*, Washington: The Brookings Institution, 1985.
- Domar, E. D., "Capital Expansion, Rate of Growth, and Employment," *Econometrica*, April 1946, 14, 137-47.
- Harrod, R. F., "An Essay on Dynamic Theory," *Economic Journal*, March 1939, 49, 14-33.
- Hickman, B., "Real Wages, Aggregate Demand, and Unemployment," *European Economic Review*, December 1987, 31, 1531-60.
- and Coen, R., *An Annual Growth Model of the U.S. Economy*, Amsterdam: North-Holland, 1976.
- Hicks, J. R., *A Contribution to the Theory of the Trade Cycle*, Oxford: Clarendon Press, 1950.
- Lewis, W. Arthur, "Economic Development with Unlimited Supplies of Labour," *The Manchester School of Economics and Social Studies*, May 1954, 22, 139-91.
- Malinvaud, E., "Notes on Growth Theory with Imperfectly Flexible Prices," in *Modern Macroeconomic Theory*, J.-P. Fitoussi, ed., Oxford: Basil Blackwell, 1983.
- Prescott, E., "Theory Ahead of Business Cycle Measurement," Working Paper, Federal Reserve Bank of Minneapolis, February 1986.
- Solow, R., "A Contribution to the Theory of Economic Growth," *Quarterly Journal of Economics*, February 1956, 70, 65-94.
- , "Technical Change and the Aggregate Production Function," *Review of Economics and Statistics*, August 1957, 39, 312-20.
- , "Investment and Technical Progress," in *Mathematical Methods in the Social Sciences, 1959*, K. Arrow, S. Karlin, and P. Suppes, eds., Stanford: Stanford University Press, 1960.
- , Tobin, J., von Weizsaecker, C. and Yaari, M., "Neoclassical Growth with Fixed Factor Proportions," *Review of Economic Studies*, April 1966, 33, 79-115.
- Wolff, E., "Capital Formation and Long-Term Productivity Growth," Working Paper, C. V. Starr Center for Applied Economics, New York University, September 1987.

# International Macroeconomic Policy Coordination When Policymakers Do Not Agree on the True Model

By JEFFREY A. FRANKEL AND KATHARINE E. ROCKETT\*

*When international policymakers do not agree on the correct macroeconomic model, they will still be able to agree on a cooperative policy package that each believes will improve welfare; but the package may turn out to move the target variables in the wrong direction. Using ten leading econometric models that could represent U.S. beliefs, non-U.S. beliefs, and the true model, we find that monetary coordination improves U.S. welfare in only 546 cases out of 1,000.*

International policy coordination is the fastest-growing research topic in the field of open-economy macroeconomics.<sup>1</sup> The topic owes its success to the happy marriage of the mathematical techniques of game theory and the practical problem of coordination that has in the mid-1980s become of central concern to international policymakers. Virtually all of the previous coordination literature has made the automatic assumption that policymakers agree on the true model of how the world macroeconomy behaves.<sup>2</sup> As a consequence, it has reached a very strong conclusion: in general, countries will be bet-

ter off if they coordinate policies than they would be in the Nash noncooperative equilibrium in which each government sets its policies while taking those of the others as given.<sup>3</sup> The empirical literature is as yet less fully developed than the theoretical literature; but it too has claimed gains from coordination that, though small, are necessarily positive.<sup>4</sup>

The assumption that policymakers agree on the true model has little, if any, empirical basis. Different governments subscribe to different economic philosophies. If one wishes to think of actors as perpetually processing new information in a Bayesian manner, so that their models over time would converge on any given reality in the limit, then one must admit that the speed of convergence is sufficiently slow, or else that reality is changing sufficiently rapidly, that policymakers

\*Department of Economics, University of California, Berkeley, CA 94720. This is a heavily revised version of NBER Working Paper No. 2059, October 1986. The authors would like to thank the Sloan Foundation and the Institute for International Studies at the University of California-Berkeley for research support, and to thank Ralph Bryant, Dale Henderson, and many seminar participants for useful comments.

<sup>1</sup>Koichi Hamada (1976) is generally credited with the birth of the topic in its modern analytic form (though under the assumption of fixed-exchange rates). More recent contributions included Matthew Canzoneri and JoAnna Gray, 1983; Marcus Miller and Mark Salmon, 1985; Kenneth Rogoff, 1985; and Willem Buiter and Richard Marston, 1985. For good introductions to the literature and further references, see Gilles Oudiz and Jeffrey Sachs, 1984; Richard Cooper, 1985; or Stanley Fischer, 1987.

<sup>2</sup>We have become aware very recently of some new papers on coordination that do allow for policymakers to be uncertain as to the true model: Atish Ghosh, 1986; Swati Ghosh and Atish Ghosh, 1986; Nouriel Roubini, 1986; Gerald Holtham, 1986; and Holtham and A. J. Hughes Hallett, 1987.

<sup>3</sup>There are two important qualifications to the generality of the proposition that coordination improves welfare under the standard assumption that policymakers know the true model. The first is that if policymakers have enough independent instruments to reach their optimum target goals regardless of one another's actions, then coordination is moot. The second is that Rogoff, 1985, and Patrick Kehoe, 1986, have shown that if coordination reduces governments' ability to precommit to anti-inflationary policies credibly to their own peoples, then it can reduce welfare. The present paper is a counterexample along very different lines.

<sup>4</sup>Oudiz and Sachs, 1984; Nicholas Carlozzi and John Taylor, 1985; Oudiz, 1985; Naoko Ishii, Warwick McKibbin, and Sachs 1985; Hughes Hallett, 1985; and Canzoneri and Patrick Minford, 1986.



have not been able to reach agreement on the true model. Nor is there much prospect of their doing so in the foreseeable future.

Professional economists are not much more able to agree on the correct macroeconomic model than policymakers. A concrete illustration was offered by a recent exercise in which a group of economists working under the auspices of the Brookings Institution asked those responsible for twelve leading econometric models of the world economy to simulate the effects of some carefully specified policy changes.<sup>5</sup> The predictions of the models varied widely as to both the magnitude and the sign of the effects on output, inflation, exchange rates, and current account balances, among trading partners and even in the country carrying out the policy change. (See Tables 1 and 8 below.) At best, no more than one of the models can be right, and it seems unlikely that even one of them is exactly right.

Lack of knowledge as to the true model helps explain a troublesome fact. While support for the proposition that coordination would improve welfare is widespread, proponents do not generally agree on the nature of the Pareto-improving package of policy changes that is called for in any particular set of circumstances. Some call for coordinated expansion, some for coordinated discipline, some for coordinated shifts in the mix between monetary and fiscal policy, and so forth.<sup>6</sup> Disagreement, even within one country, as to where the economy currently sits relative to the desired values of the target variables is responsible for some of the disagreement on the desirable coordinated policy changes, but disagreement as to the correct model is also a significant factor. As William Branson (1986, p. 176) says, "With

this range of disagreement on economic analysis, how are the negotiators to reach agreement? The topic is one for the National Science Foundation, not a new Bretton Woods." Martin Feldstein (1983) argues similarly.

One implication of the lack of agreement on the true model is, of course, that "more research needs to be done." But the implications for any policy coordination that might take place in the meantime are considerably more interesting than this familiar platitude. This paper makes three points relevant when policymakers disagree on the model. First, such policymakers will in general be able to find a package of coordinated policy changes that each believes will improve its country's welfare relative to the suboptimal Nash non-cooperative equilibrium.<sup>7</sup> Second, and in striking contrast to the standard result when policymakers know the true model, the package of coordinated policy changes often turns out to reduce welfare, as judged by some true model of reality, rather than to raise it. For example, using ten models from the Brookings' simulations as models which could represent the views of the U.S. government, the views of other industrialized countries, or the true world macroeconomy, we find that out of 1,000 possible combinations, monetary coordination perceptibly improves U.S. welfare in only 546 cases, and improves the welfare of the other industrialized countries in only 539 cases. Third, the gains to one country from unilaterally discovering the true model and adjusting its policy accordingly, are usually much greater than the potential gains from coordination. We find, as have others, that the gains from coordination turn out to be very small even when the odds are stacked in its favor by assuming that both parties know the true model.

<sup>5</sup>The project was entitled "Empirical Macroeconomics for Interdependent Economies," and is forthcoming as Bryant et al., 1988. Jeffrey Frankel (1986a) discusses the disagreements among the 12 models.

<sup>6</sup>Some of the authors in the coordination literature decline to take any position at all on whether the problem with the Nash noncooperative equilibrium is that it is too contractionary or too expansionary, etc. They leave it for econometricians to fill in the correct parameter values at some later date.

<sup>7</sup>One's intuition is that players who disagree about the model will find it harder to agree on a package of joint policy changes (for example, Richard Cooper, 1986). The correct way to interpret this intuition is probably that, even if there exists a bargaining solution that is believed to be Pareto-superior to the noncooperative solution, it will be harder for the players to agree on a mechanism to enforce the bargaining solution if they do not share a common view of the world.

Sections I and II of the paper analyze a very simple game where two countries, the United States and "Europe" (shorthand for the non-U.S. OECD) must decide how to set their money supplies so as to come as close as possible to their desired levels of two target variables: income and the current account (internal balance and external balance). Section I makes the two points theoretically, that the two central banks will in general be able to agree on a coordinated policy package that each thinks leaves its country in a better position, and that the package might in fact leave them in a worse position. Section II uses the multipliers from the ten models in the Brookings' simulation to provide a dramatic illustration of the points.

In Section III each government is given a second policy instrument, government expenditure, to use in addition to monetary policy, and a third target variable, inflation, to pursue in addition to income and the current account. Again we see that the governments will in general find a coordinated policy package that they expect to improve welfare, but that it often has the opposite effect in reality. We conclude in Section IV by mentioning extensions of the framework to deal with the policymaker's uncertainty regarding the true model, or the other player's model, or both.

### I. The Theory of Monetary Coordination with Disagreement

Here we assume that each country is interested in two target variables: its own output, denoted  $y$  for the United States and  $y^*$  for Europe (expressed in log form and relative to their optimum values), and its current account balance, denoted  $x$  and  $x^*$ , respectively (expressed as a percentage of GNP and again relative to their optima). Each government seeks to minimize a quadratic loss function.

$$(1) \quad W = y^2 + \omega x^2$$

$$(2) \quad W^* = y^{*2} + \omega^* x^{*2}$$

where  $\omega$  and  $\omega^*$  denote the relative weights

placed on external balance versus internal balance.

We assume a general framework in which the targets are linearly related to the available policy instruments, which in this section are limited to the countries' money supplies,  $m$  and  $m^*$ , respectively (in log form). We denote the parameters as perceived by the U.S. authorities by a "us" subscript.

$$(3) \quad y = A_{us} + C_{us}m + E_{us}m^*$$

$$(4) \quad x = B_{us} + D_{us}m + F_{us}m^*$$

We denote the parameters perceived by the European government by an "e" subscript.

$$(5) \quad y^* = G_e + I_e m + K_e m^*$$

$$(6) \quad x^* = H_e + J_e m + L_e m^*$$

Since each country has only a single instrument but two targets, it cannot unilaterally achieve its targets. We begin by considering the Nash noncooperative equilibrium. To ascertain U.S. behavior, we differentiate (1) with respect to  $m$ , using (3) and (4) and holding  $m^*$  constant. It follows that the U.S. reaction function is

$$(7) \quad m = M + N m^*$$

$$\text{where } M = - \frac{A_{us}C_{us} + \omega B_{us}D_{us}}{C_{us}^2 + \omega D_{us}^2},$$

$$\text{and } N = - \frac{E_{us}C_{us} + \omega F_{us}D_{us}}{C_{us}^2 + \omega D_{us}^2}.$$

To ascertain European behavior we differentiate (2) with respect to  $m^*$ , using (5) and (6) and holding  $m$  constant. The European reaction function is

$$(8) \quad m^* = Q + R m,$$

$$\text{where } Q = - \frac{G_e K_e + \omega^* H_e L_e}{K_e^2 + \omega^* L_e^2}$$

$$\text{and } R = - \frac{I_e K_e + \omega^* J_e L_e}{K_e^2 + \omega^* L_e^2}.$$

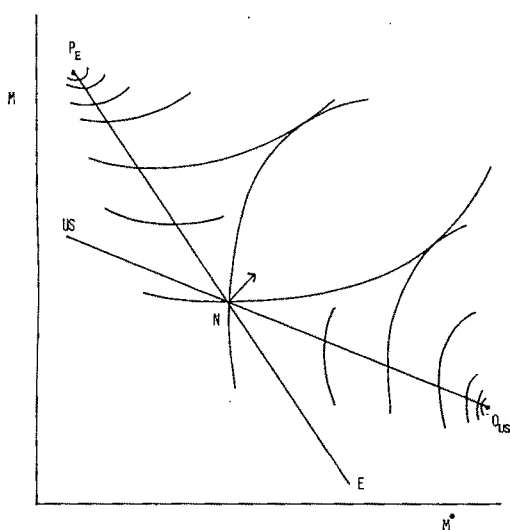


FIGURE 1

We solve equations (7) and (8) for the Nash equilibrium.

$$(9) \quad m^n = \frac{M + NQ}{1 - NR},$$

$$(10) \quad m^{*n} = \frac{Q + MR}{1 - NR}.$$

Figure 1 shows the two policymakers' reaction functions, equations (7) and (8). The optimum point as perceived by the U.S. policymakers is a point  $0_{us}$  on its reaction function. Concentric indifference curves radiate from  $0_{us}$ . These curves are vertical wherever they intersect the reaction function, because  $m$  is chosen so that its marginal benefit given  $m^*$  is zero. Similarly the optimum point as perceived by the European policymaker is a point  $P_e$ , and its concentric indifference curves are horizontal wherever they intersect its reaction function.

We have drawn the European reaction curve as steeper than the U.S. curve. One might expect the effects that are largest in absolute value to be the positive effects of money on domestic output:  $C$  in equations (3)–(4) for the United States and  $K$  in equa-

tions (5)–(6) for the non-U.S. OECD.<sup>8</sup> It would follow that, unless the welfare weight  $\omega$  on the current account is large, the absolute value of the slope of the U.S. reaction function is less than one when the U.S. money supply is on the vertical axis, and vice versa for the European reaction function.

The possibilities for the sign of the slope are more diverse. If monetary expansion is thought to be transmitted negatively to trading partners ( $E < 0$ ), presumably via a depreciation of the currency and improvement in the trade balance of the expanding country as in the Mundell-Fleming model, then the slope is positive:  $N > 0$ . If monetary transmission is thought to be positive on the other hand ( $E > 0$ ), then the slope is ambiguous: when the welfare weight  $\omega$  on the current account is small, the slope is negative, but when  $\omega$  is large, or when the transmission multiplier  $E$  is small (relative not only to the own multiplier  $C$ , but also to the current account multipliers  $D$  and  $F$ ), the slope is again positive. (We are assuming that  $D$  and  $F$ , the effects of  $m$  and  $m^*$  on the domestic current account, are of opposite signs by symmetry.)

The same analysis holds for the foreign reaction function (for example,  $I < 0 \Rightarrow R > 0$ ), though it must be remembered that even if any given model is symmetric, the two reaction functions could easily have opposite slopes. For example one country might believe that transmission is negative and the other that it is positive. In Figure 1 we have drawn the functions downward sloping: a foreign expansion is transmitted positively to the domestic country and so the domestic government reacts by contracting.

The Nash equilibrium  $N$  is determined as the intersection of the two reaction functions. At  $N$  the indifference curves cannot be tangent, but must intersect, since their respective slopes are infinity and zero. It follows that the Nash equilibrium is per-

<sup>8</sup>This holds in the ten econometric models considered in the following section except the LIVPL and MSG models for the U.S. and LIVPL, MSG, Wharton, and EPA models for Europe.

ceived as Pareto-inefficient. Both policy-makers think they would be better off if they could agree to move to a point within the "lens" determined by the intersection of the two indifference curves.

As we have drawn the graph, each country would like to expand but is afraid to do so on its own, presumably because of adverse implications for the current account. But they can agree to expand simultaneously, moving northeastward in the graph to higher levels of perceived welfare. Such joint reflation is the kind of international coordination that has been urged on Germany and Japan by the United States under two different administrations: in 1977-78, in the form of the "locomotive theory," and in 1986 in the form of coordinated discount rate cuts.<sup>9</sup>

If an efficient mechanism of coordination exists, the countries will move, not just northeastward, but specifically to one of the points on the contract curve, where the two countries' indifference curves are tangent. There is no strong reason to choose any particular point. Nor, for that matter, is there reason to think that any Pareto-improving solution can necessarily be enforced. But we follow much of the literature in considering the Nash-bargaining solution, defined as the point where the product of the two countries' perceived welfare gains, compared to the perceived welfare at the Nash noncooperative solution, is maximized.<sup>10</sup>

$$(11) \quad \text{Max} (W_{us}(m, m^*) - W_{us}(m^n, m^{*n})) \\ \times (W_e^*(m, m^*) - W_e^*(m^n, m^{*n}))$$

subject to equations (1)-(6).

One would differentiate with respect to  $m$  and  $m^*$  to find the bargaining solution ( $m^b, m^{*b}$ ), a point such as  $B$  in Figure 2.

<sup>9</sup>More often, it has been private economists, and the governments of smaller countries, who have urged such coordinated expansion; for example, Fred Bergsten et al. (1982). The 1981-84 Reagan Administration opposed coordination.

<sup>10</sup>In a related exercise, Holtham and Hughes Hallett, 1987, p. 26, show that the results change little when definitions of the bargaining equilibrium other than the Nash solution are used.

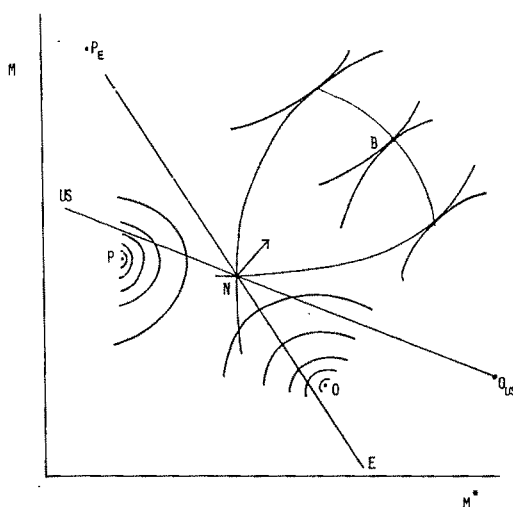


FIGURE 2

Once we recognize that the two policy-makers have different models of the world, we must recognize that one, or both, will be wrong. To evaluate whether the bargaining solution  $B$  is superior to the noncooperative solution ( $m^n, m^{*n}$ ) not just in perception but also in reality; we would have to know the true parameter values, the output and current account functions (3)-(6) without the subscripts:

$$(12) \quad y = A + Cm + Em^*$$

$$(13) \quad x = B + Dm + Fm^*$$

$$(14) \quad y^* = G + Im + Km^*$$

$$(15) \quad x^* = H + Jm + Lm^*.$$

We would then plug  $m^b$  and  $m^{*b}$  into (12)-(15), and in turn plug the target variables into the loss functions (1) and (2), to see whether the bargaining solution in fact improves welfare.

In the standard case where the policy-makers both know the correct model, coordination must necessarily improve welfare for each country, or else its government would not have agreed to go along. In our case, coordination *may* improve welfare. For example, if the true model is very close to that

believed by the U.S. authorities, then the true iso-welfare map will be very similar to the perceived indifference curves shown in Figure 1, and U.S. welfare will indeed be higher at *B* than *N*. But this need not be the case.

The true optimum policy combination to maximize U.S. welfare is given by differentiating (1) with respect to *m* (as in the derivation of (7) but without the subscripts), and with respect to *m*<sup>\*</sup>, and solving simultaneously:

$$(16) \quad m^0 = \frac{M(E^2 + \omega F^2) - N(AE + \omega BF)}{(E^2 + \omega F^2) + N(CE + \omega DF)},$$

$$(17) \quad m^{*0} = -\frac{AE + \omega BF}{E^2 + \omega F^2} - \frac{CE + \omega DF}{E^2 + \omega F^2} m^0.$$

If the true optimum point 0 is not at 0<sub>us</sub> but rather is as shown in Figure 2, with the new set of true iso-welfare curves drawn, then the move from *N* to *B* could very well be in the wrong direction, resulting in a reduction in U.S. welfare. Similarly if the true optimum policy combination from the viewpoint of European interests is not at *P*<sub>e</sub> but rather at *P* as shown in Figure 2, then coordination could reduce European welfare as well.

It is worth considering momentarily the case when the two policymakers are seeking to maximize the identical objective function, and disagree *only* about the proper model. For example they might be the monetary and fiscal authority within the same country. Our two propositions would still hold: (1) the two policymakers will in general be able to agree on a package of coordinated policy changes that each thinks will improve the (same) country's welfare relative to the Nash noncooperative solution, and (2) the package agreed to in bargaining could in fact worsen welfare as easily as improve it. This is the case considered in Frankel (1986b).<sup>11</sup> While in that paper conflict and coordination arise

solely from different perceptions, and in the conventional literature they arise solely from different objectives, in the present paper both factors are present.

## II. Coordination with Ten International Econometric Models

How important for coordination is the issue of conflicting models likely to be in practice? Is the case in which bargaining reduces welfare as judged by the true model merely a pathological counterexample, or is it a likely occurrence? In what follows we use the international simulation results of the macroeconomic models that participated in the Brookings' exercise to get an idea of what might actually happen if governments coordinate.

The models were asked to show the effects of four experiments, among others: an increase in the U.S. money supply, an increase in the non-U.S. OECD money supply, an increase in U.S. government expenditure, and an increase in non-U.S. OECD government expenditure. In each case the instructions were to hold the other policy instruments constant. Though twelve models participated, some did not report effects on current account balances, which we need along with effects on output levels. The ten that we can use here are the Federal Reserve Board's Multi-Country Model (MCM), Patrick Minford's Liverpool Model (LIVPL), the Sims-Litterman Vector Auto-Regression Model (VAR), the OECD's Interlink Model (OECD), the Project Link Model (LINK), the McKibbin-Sachs Global Model (MSG), the EEC Commission's Compact Model (EEC), the Haas-Masson smaller approximation of the MCM model (MINIMOD), the Economic Planning Agency model (EPA), and the Wharton model (Wharton). These models are quite representative of the range of econometric models actually in use, including as they do models both large and small in size, structural and nonstructural in approach, Keynesian and neoclassical in philosophy, backward-looking and forward-looking in expectations formation, public-sector and private-sector in function, and non-American and American in authorship.

<sup>11</sup>In equations (3) and (4), one could simply redefine *m*<sup>\*</sup> as fiscal policy, and let *y*<sup>\*</sup> ≡ *y*, *x*<sup>\*</sup> ≡ *x*, and *ω*<sup>\*</sup> ≡ *ω*. As long as the two policymakers have different parameter estimates, there will still be scope for coordination. One difference is that in Figure 2 the true optimal points *P* and *O* would coincide.

TABLE 1—MONETARY POLICY SIMULATION EFFECT IN SECOND YEAR OF INCREASE IN MONEY SUPPLY (4 PERCENT)<sup>a,b</sup>

	Y	CPI	i	Currency Value	CA	CA*	i*	CPI*	Y*
Monetary Expansion in United States	Effect in United States				Effect in Rest of OECD				
	(in percent)	(in percent)	(Pts.)	(in percent)	(\$b)	(\$b)	(Pts.)	(in percent)	
MCM	+1.5	+0.4	-2.2	-6.0	-3.1	-3.5	-0.5	-0.6	-0.7
EEC <sup>c</sup>	+1.0	+0.8	-2.4	-4.0	-2.8	+1.2	-0.5	-0.4	+0.2
EPA <sup>d</sup>	+1.2	+1.0	-2.2	-6.4	-1.6	-10.1	-0.6	-0.5	-0.4
LINK	+1.0	-0.4	-1.4	-2.3	-5.9	+1.5	NA	-0.1	+0.1
LIVERPOOL	+0.1	+3.7	-0.3	-3.9	-13.0	+0.1	-0.1	-0.0	-0.0
MSG	+0.3	+1.5	-0.8	-2.0	+2.6	-4.4	-1.2	-0.7	+0.4
MINIMOD	+1.0	+0.8	-1.8	-5.7	+2.8	-4.7	-0.1	-0.2	-0.2
VAR <sup>e</sup>	+3.0	+0.4	-1.9	-22.9	+4.9	+5.1	+0.3	+0.1	+0.4
OECD	+1.6	+0.7	-0.8	-2.6	-8.4	+3.1	-0.1	-0.1	+0.3
TAYLOR <sup>e</sup>	+0.6	+1.2	-0.4	-4.9	NA	NA	-0.1	-0.2	-0.2
WHARTON	+0.7	+0.0	-2.1	-1.0	-5.1	+5.3	-1.3	-0.1	+0.4
DRI	+1.8	+0.4	-2.3	-14.6	-1.4	+14.5	-1.1	-1.3	-0.6

	Y	CPI	i	Currency Value	CA	CA*	i	CPI*	Y*
Monetary Expansion in Rest of OECD	Effect in Rest of OECD				Effect in United States				
	(in percent)	(in percent)	(Pts.)	(in percent)	(\$b)	(\$b)	(Pts.)	(in percent)	
MCM	+1.5	+0.6	-2.1	-5.4	+3.5	+0.1	-0.2	-0.2	-0.0
EEC <sup>c</sup>	+0.8	+1.0	-1.0	-2.3	-5.2	+1.9	+0.0	+0.1	+0.1
EPA <sup>d</sup>	+0.0	+0.0	-0.1	-0.1	-0.1	+0.1	-0.0	-0.0	+0.0
LINK <sup>f</sup>	+0.8	-0.6	NA	-2.3	-1.4	+3.5	+0.0	-0.0	+0.1
LIVERPOOL	+0.4	+2.8	-0.9	-8.4	+7.1	-8.2	-1.1	-3.4	+1.6
MSG	+0.2	+1.5	-0.7	-1.4	-15.9	+12.0	-1.2	-0.6	+0.3
MINIMOD	+0.8	+0.2	-1.8	-4.8	+3.6	-1.4	-0.6	-0.5	-0.3
VAR <sup>e</sup>	+0.7	-0.5	-3.0	-5.5	+5.2	-10.0	+0.6	-0.7	+1.2
OECD	+0.8	+0.3	-1.3	-2.1	-1.6	+2.3	-0.2	+0.1	+0.1
TAYLOR <sup>e</sup>	+0.8	+0.7	-0.3	-3.5	NA	NA	-0.2	-0.5	-0.1
WHARTON	+0.2	-0.1	-0.8	+0.2	+2.6	+0.5	+0.0	+0.0	+0.0
DRI	NA	NA	NA	NA	NA	NA	NA	NA	NA

<sup>a</sup> The increase in the money supply is phased in over four quarters.<sup>b</sup> Source: Frankel (1986a).<sup>c</sup> Non-U.S. short-term interest rate NA; long-term reported instead.<sup>d</sup> Non-U.S. current account is Japan, Germany, United Kingdom, and Canada.<sup>e</sup> CPI NA. GNP deflator reported instead.<sup>f</sup> Appreciation of non-U.S. currency NA; depreciation of dollar reported instead.

Table 1 reports the effects of monetary expansion on several macroeconomic variables according to each of the twelve models. The simulations showed effects over six years, but ours is a static framework; we use only the effect in the second year. The models all agree that a monetary expansion raises domestic output, but they agree on little else. There is a surprising amount of disagreement, in particular, on whether a monetary expansion improves or worsens the current account and, in turn, on whether it is trans-

mitted negatively or positively to the rest of the world. The reasons for this and other disagreements in the simulations are examined elsewhere.<sup>12</sup> It suffices to repeat that

<sup>12</sup> The positive effect of a monetary expansion on the current account via currency depreciation is offset by a negative effect via higher income. In the Mundell-Fleming model the positive effect on the current account must dominate, to match the net capital outflow that results from lower interest rates, giving negative trans-

TABLE 2—MONEY AND FISCAL MULTIPLIERS  
(FOR THREE TARGETS IN EACH COUNTRY)

	Percentage Effect on Income		Effect on Current Account (As Per- centage of GNP)		Effect on Percentage Inflation Rate	
From a (1 percent) increase in:	U.S. M	Eur. M	U.S. M	Eur. M	U.S. M	Eur. M
Effect on United States						
MCM	0.3750	0.0000	-0.0198	0.0006	0.1000	-0.0500
VAR	0.7500	0.3000	0.0311	-0.0634	0.1000	-0.1750
OECD	0.4000	0.0250	-0.0537	0.0147	0.1750	-0.0250
LINK	0.2500	0.0250	-0.0380	0.0225	-0.1000	0.0000
Effect on Europe						
MCM	-0.1750	0.3750	-0.0090	0.0090	-0.1500	0.1500
VAR	0.1000	0.1750	0.1169	0.1192	0.0250	-0.1250
OECD	0.0750	0.2000	0.0178	-0.0091	-0.0250	0.0750
LINK	-0.0250	0.2000	0.0083	-0.0077	-0.0250	-0.1500
From an increase (equal to 1 percent of GNP):	U.S. G	Eur. G	U.S. G	Eur. G	U.S. G	Eur. G
Effect on United States						
MCM	1.8000	0.5000	-0.4217	0.2019	0.4000	0.2000
VAR	0.4000	0.3000	-0.0127	-0.0659	-0.9000	-0.1000
OECD	1.1000	0.1000	-0.3628	0.0843	0.6000	0.2000
LINK	1.2000	0.2000	-0.1647	-0.1621	0.5000	0.0000
Effect on Europe						
MCM	0.7000	1.4000	0.0912	-0.0737	0.4000	0.3000
VAR	-0.0000	0.5000	-0.0183	0.1559	0.0000	-0.3000
OECD	0.4000	1.5000	0.2583	-0.1564	0.3000	0.7000
LINK	0.1000	1.2000	0.0420	-0.1349	0.0000	0.1000

disagreements with respect to both the sign and magnitude of effects are common among honorable economists, and are common even within subsets of models that are supposedly similar in orientation, let alone among policymakers.

The first half of Table 2 reports multipliers for output and the current account calculated in the form that we need: as a percentage of GNP per one percent change in the money supply. To save space in this and other tables below, we report numbers

for only four of the models: MCM, VAR, OECD, and LINK. Numbers are reported for six models, including also the LIVPL and MSG models, in the tables in the NBER Working paper version of this study. The qualitative outcome of calculations will be reported here for all ten models, including also the EEC, MINIMOD, Wharton, and EPA models.

Computing the policymakers' reactions requires knowing not only the perceived policy multipliers, but also the target optima and the welfare weights. We adopt the same target values as Oudiz and Sachs (1984): current accounts of zero for the United States and two percent of GNP for the non-U.S. OECD, and GNP gaps of zero for both regions. The baseline values of both variables, specified as part of the Brookings' simulation exercise, were below target as of 1985. Thus policymakers will seek to increase both output and the current account. The targets, together with the baseline values for the variables and

mission abroad. But in more modern models the net capital flow may be reversed, in response to perceived overshooting of the exchange rate. The theoretical literature contains many other ways of reversing the Mundell-Fleming transmission results as well. (See Michael Mussa, 1979, or, for an optimizing approach, Lars Svensson and Sweder van Wijnbergen, 1986.) On the models used in the Brookings' simulations, see Frankel, 1986a, or other papers in Ralph Bryant et al., 1988.

any set of policy multipliers from Table 2, imply corresponding values for the constant terms  $A$ ,  $B$ ,  $G$ , and  $H$  in equations (3)–(6).<sup>13</sup>

The choice of welfare weights  $\omega$  and  $\omega^*$  is necessarily more arbitrary, even, than the choice of target optima. Oudiz and Sachs chose the values that the weights would have had to have held for countries to have produced the values of output, inflation, and the current account actually observed in the 1980s, assuming a Nash noncooperative equilibrium. For lack of a better alternative, we adopt the set of weights calculated by Oudiz and Sachs for the EPA model, and apply it uniformly regardless of model. It would be of questionable benefit to replicate their methodology separately with each model; our welfare comparisons require a common objective function. On the other hand, setting a common set of weights for all models has the drawback that the Nash solution may lie very far from the baseline for certain combinations of models. Often, in our simulations, such large moves to the Nash point resulted in a loss in welfare which was recovered in a similarly large move to the coordination point. In general, our experiment may bias our results a little toward gains from coordination because often all the models agreed on the move from the remote noncooperative point. In

order to test the sensitivity of our results to the chosen weights, we repeated the experiment two more times, using different weights. The alternative weights were obtained by using Oudiz and Sachs' methodology, pegging the Nash equilibrium to the baseline, for two models: the OECD and LIVPL models, respectively. The odds in favor of coordination changed little.<sup>14</sup>

If the U.S. policymaker can believe any of the ten models and the non-U.S. (henceforth "European") policymaker can believe any of the ten models, then there are  $10 \times 10 = 100$  possible combinations, each implying a different Nash noncooperative equilibrium. For each combination we computed the values of the two countries' variables of interest in the Nash noncooperative equilibrium: the money supply, the perceived output and current account, and the perceived welfare function. All but one of the sixteen cases we report called for expansion from the baseline by one country or the other.<sup>15</sup>

Our main interest lies in the move from the noncooperative to the bargaining equilibrium, shown in Table 3. To take one example, if the U.S. policymaker believes in the MCM model and the European policymaker believes in the OECD model, then

<sup>13</sup>Our objective functions (1) and (2), like those of Oudiz and Sachs, and others, pertain only to welfare in the current period. They thus neglect any dynamic effects such as the effect current expansion has on next period's inflationary expectations and therefore on next period's welfare. Propositions stated here about the desirability of changes in the levels of policy variables could be reinterpreted as propositions about the desirability of changes in policy rules, using dynamic models with rational expectations. See Taylor (1985). But the question of whether governments should coordinate policies is usually in practice intended to refer to regular meetings in which representatives discuss the current setting of their money supplies and other variables rather than a one-time "global constitutional convention" in which they discuss whether to adopt strict monetarist rules versus, for example, nominal income targeting. Most of the ten models used here to represent policymakers' beliefs, with the exception of the MSG and Taylor models, are designed to predict the effects of changes in policy variables rather than changes in policy rules.

<sup>14</sup>After testing for sensitivity to the choice of weights, we also tested for sensitivity to the choice of targets. The total count for true gains and losses for the two countries were:

		U.S.	Europe
OECD weights;	gains	507	479
	losses	322	349
	zeroes	171	172
$\omega = 1/11.8$ $\omega^* = 1/244.0$	gains	421	471
	losses	302	267
	zeroes	277	262
LIVPL weights; $\omega = 1/.26$ $\omega^* = 1/2.4$	gains	538	537
	losses	338	340
	zeroes	124	123
Original weights; target level of U.S. GNP = 95 percent of baseline	gains	484	465
	losses	272	291
	zeroes	244	244

<sup>15</sup>The numbers for  $6 \times 6 = 36$  combinations are reported in the NBER paper.



TABLE 3—THE COOPERATIVE BARGAIN<sup>a</sup>

Model Subscribed to by the United States	Model Subscribed to by Europe			
	<i>MCM</i>	<i>VAR</i>	<i>OECD</i>	<i>LINK</i>
<i>MCM</i>				
Bargaining Change in Policy				
Meur <sup>b</sup>	0.240	2.020	1.590	0.710
Mus	-0.137	-0.441	0.367	0.253
Perceived Change in Targets				
Europe	<i>Y</i> 0.114	0.309	0.346	0.136
	<i>CA</i> 0.003	0.189	-0.008	-0.003
United States	<i>Y</i> -0.051	-0.165	0.138	0.095
	<i>CA</i> 0.003	0.010	-0.006	-0.005
Perceived Gain for				
Europe	0.0001	0.0066	0.0011	0.0002
United States	0.0000	0.0002	0.0002	0.0001
<i>VAR</i>				
Bargaining Change in Policy				
Meur	-5.199	25.185	-32.772	-4.129
Mus	-8.652	-17.173	32.898	8.394
Perceived Change in Targets				
Europe	<i>Y</i> -0.436	2.690	-4.087	-1.036
	<i>CA</i> 0.031	0.995	0.875	0.102
United States	<i>Y</i> -8.049	-5.325	14.842	5.057
	<i>CA</i> 0.061	-2.130	3.099	0.522
Perceived Gain for				
Europe	0.0092	0.3216	0.2415	0.0090
United States	0.3333	0.4323	3.2174	0.4100
<i>OECD</i>				
Bargaining Change in Policy				
Meur	0.389	14.490	3.837	2.142
Mus	-0.533	-8.940	2.356	1.820
Perceived Change in Targets				
Europe	<i>Y</i> 0.239	1.642	0.944	0.383
	<i>CA</i> 0.008	0.682	0.007	-0.001
United States	<i>Y</i> -0.204	-3.214	1.038	0.782
	<i>CA</i> 0.034	0.693	-0.070	-0.066
Perceived Gain for				
Europe	0.0004	0.1611	0.0078	0.0015
United States	0.0010	0.0933	0.0128	0.0064
<i>LINK</i>				
Bargaining Change in Policy				
Meur	0.388	35.304	4.975	3.479
Mus	-0.851	-29.058	5.112	4.525
Perceived Change in Targets				
Europe	<i>Y</i> 0.294	3.272	1.378	0.583
	<i>CA</i> 0.011	0.811	0.045	0.011
United States	<i>Y</i> -0.203	-6.382	1.402	1.218
	<i>CA</i> 0.041	1.898	-0.082	-0.093
Perceived Gain for				
Europe	0.0006	0.6101	0.0161	0.0038
United States	0.0010	0.3590	0.0251	0.0141

<sup>a</sup>For this and all subsequent tables, changes in instruments and targets are expressed in percent, and changes in utility are expressed in percent-squared GNP.

<sup>b</sup>For this and all subsequent tables, the "eur" subscript refers to European variables, and the "us" subscript refers to the United States' variables.

they can agree to expand further their money supplies simultaneously (0.37 percent and 1.59 percent, respectively). They each believe that this policy package will result in higher output with little adverse effect on their current accounts. This is the often-mentioned case in which the Nash equilibrium is too contractionary. But besides the case of simultaneous expansion (six combinations of models in this table), every other case is possible as well: European expansion with U.S. contraction (7 combinations), U.S. expansion with European contraction (2 combinations), and simultaneous contraction (1 combination).

Without knowing the true model, we cannot determine whether any given policy package actually improves welfare. But we can get a good idea of the possibilities by trying out each of the models as a candidate for the true model. The sixteen cells in Tables 4 and 5 correspond to the same sixteen combinations as Table 3. But within each cell we report the effect that the corresponding coordination package of Table 3 would have under each of the 4 models; thus there are  $4^3 = 64$  combinations in all.<sup>16</sup> Table 4 shows the actual effect of coordination on U.S. welfare and Table 5 the effect on European welfare. Whenever one or the other policymaker turns out to have had the right model, his country does gain from coordination. Otherwise he would not have agreed to the package. For example the joint monetary expansion that they agree on when the U.S. policymaker believes the MCM model and the European policymaker believes the OECD model is seen to raise U.S. welfare if the MCM model is the true one (Table 4) and to raise European welfare if the OECD model is the true one (Table 5). It also turns out to raise both countries' welfare if the LINK model is the true one. But it turns out to *reduce* welfare if the VAR model is the correct one (also the LIVPL and MSG mod-

els). The reader who does not believe in one of the latter three models might not be concerned with that result. But such a reader should instead be concerned with the result that when the U.S. policymaker, for example, believes in the LIVPL model and the European policymaker in the VAR model, coordination will reduce welfare according to each of the other models.<sup>17</sup>

It must be noted that it is the countries' failure to perceive the true model, not their failure to agree with each other per se, that alters the standard conclusion regarding coordination. The case for coordination is not necessarily any stronger if the countries agree on one model and it turns out to be the wrong model. On the other hand, the point in this paper is something stronger than simply "bad models lead to bad policies." In domestic policymaking, we are often fairly confident that we know the *sign* of policy effects and are uncertain only about the magnitude. Though it follows from such uncertainty that policymakers should be more timid about policy changes, there will still at least be some small change that will move the economy closer to its desired targets (William Brainard, 1967). But in international policymaking, we are uncertain even as to the sign of policy effects, such as monetary transmission between countries. For this reason, even a small step in the direction of

<sup>16</sup>The diagonal entries of the three-dimensional matrix are the cases in which both policymakers have the correct model. These calculations correspond conceptually to those in Oudiz and Sachs (1984) for the MCM and EPA models.

<sup>17</sup>The most bizarre combination occurs when the U.S. believes the LIVPL model and Europe believes the OECD model (not shown). Under this combination, the Nash noncooperative equilibrium entails a mutually destructive increase in the European money supply of almost 100 percent and decrease in the U.S. money supply of over 100 percent (!). Evidently the problem is that the LIVPL model shows European monetary expansion raising U.S. output much more than does U.S. monetary expansion, as can be seen in Table 1. There is no reason why the Nash solution for the money supply specified in equations (9) and (10) need be positive. One need only plug in the multiplier values from Table 2 to see how negative money supplies are possible. Presumably, in practice, U.S. policymakers would begin to doubt the LIVPL model and its prediction that European monetary expansion would have such a powerful expansionary effect on U.S. output, long before they relied on it to the extent of reducing the U.S. money supply to zero.

TABLE 4—TRUE GAINS FROM COORDINATION FOR THE UNITED STATES

Model Subscribed to by the United States	Model Subscribed to by Europe			
	<i>MCM</i>	<i>VAR</i>	<i>OECD</i>	<i>LINK</i>
<i>MCM</i>				
Model Representing Reality				
<i>MCM</i>	0.0000	0.0002	0.0002	0.0001
<i>VAR</i>	0.0001	0.1068	-1.0427	-0.5630
<i>OECD</i>	0.0047	0.0313	0.0004	-0.0024
<i>LINK</i>	0.0019	0.0596	0.0119	0.0041
<i>VAR</i>				
Model Representing Reality				
<i>MCM</i>	-2.1398	1.0677	6.0012	1.5986
<i>VAR</i>	0.3333	0.4323	3.2174	0.4100
<i>OECD</i>	-2.5239	1.6538	6.3684	1.7626
<i>LINK</i>	-1.0564	1.3516	2.4973	0.7067
<i>OECD</i>				
Model Representing Reality				
<i>MCM</i>	-0.0127	-0.1063	0.0362	0.0301
<i>VAR</i>	0.2994	-0.6632	-3.5351	-2.4221
<i>OECD</i>	0.0010	0.0933	0.0128	0.0064
<i>LINK</i>	-0.0008	0.3912	0.0357	0.0203
<i>LINK</i>				
Model Representing Reality				
<i>MCM</i>	-0.0138	-1.4439	0.0337	0.0341
<i>VAR</i>	0.6384	-5.2320	-6.3935	-5.3937
<i>OECD</i>	0.0088	-1.1929	-0.0491	-0.0470
<i>LINK</i>	0.0010	0.3590	0.0251	0.0141

TABLE 5—TRUE GAINS FROM COORDINATION FOR EUROPE

Model Subscribed to by the United States	Model Subscribed to by Europe			
	<i>MCM</i>	<i>VAR</i>	<i>OECD</i>	<i>LINK</i>
<i>MCM</i>				
Model Representing Reality				
<i>MCM</i>	0.0001	1.9596	0.2717	0.0974
<i>VAR</i>	-0.1516	0.0066	-2.0826	-1.0593
<i>OECD</i>	-0.0120	0.3715	0.0011	-0.0050
<i>LINK</i>	-0.0127	0.4272	0.0119	0.0002
<i>VAR</i>				
Model Representing Reality				
<i>MCM</i>	0.0092	31.7840	-4.1226	-0.9188
<i>VAR</i>	11.2505	0.3216	1.2862	-3.0464
<i>OECD</i>	-0.1383	4.0254	0.2415	0.0886
<i>LINK</i>	0.0729	6.0950	0.3258	0.0090
<i>OECD</i>				
Model Representing Reality				
<i>MCM</i>	0.0004	16.0344	0.4762	0.2004
<i>VAR</i>	0.1519	0.1611	-6.4749	-4.2203
<i>OECD</i>	-0.0153	2.2176	0.0078	-0.0054
<i>LINK</i>	-0.0222	3.1398	0.0185	0.0015
<i>LINK</i>				
Model Representing Reality				
<i>MCM</i>	0.0006	39.3135	0.4532	0.2078
<i>VAR</i>	0.5566	0.6101	-10.6004	-8.6072
<i>OECD</i>	-0.0124	4.6683	0.0161	-0.0104
<i>LINK</i>	-0.0252	7.3543	0.0271	0.0038

a coordinated policy change may lower welfare.

Altogether there are  $10^3 = 1000$  combinations counting those not shown in the tables. Coordination turns out to result in gains for the United States in 546 cases, as against losses in 321 cases and no perceptible effect (to four decimal places) in 133 cases. For Europe there are gains in 539 cases, as against losses in 327 cases and no effect in 134 cases. These figures in a sense overstate the odds in favor of successful coordination, in that by construction each country's welfare is improved (or at least not worsened) in 1/10 of the combinations, those in which the policymaker has the same model as the true one. If we take only the  $10 \times 9 \times 9 = 810$  combinations where neither country is correct, the proportion of losses is higher. For the United States there are gains in 419 cases, as against losses in 286 cases and no effect in 105. For Europe there are gains in 408 cases, losses in 298, and no effect in 104.

The results thus suggest that the danger that coordination will worsen welfare rather than improve it is more than just a pathological counterexample. One cannot, under conditions where policymakers subscribe to different models, make the blanket pronouncement that coordination as it is conventionally defined must improve welfare.

It would be helpful to know whether the incidence of gains vs. losses from coordination can be associated with any particular pattern in the policymakers' perceptions. Such knowledge might allow us to devise alternative concepts of cooperation that would be more likely to improve welfare.

In our framework, losses may occur because countries make errors on the sign of the multipliers, so they adjust their instruments in the wrong direction from the true utility-improving direction. On the other hand, losses may also occur if countries are correct about the sign of the multipliers, but perceive monetary policy to be less effective than it is in reality. In this case, they may adjust their instruments in the direction of the true coordination point, but adjust the instruments too much. This may result in a loss in welfare if "overshooting" is severe. A simple remedy for the second type of loss

would be to make smaller moves to reduce losses due to the incorrect magnitude effect, leaving only sign errors. In our simulations, overshooting turns out to be the cause of the losses from coordination in only 25 out of the 189 cases in which coordination results in losses for both countries. The primary reason for losses in our simulations is, then, moves in the wrong direction. Given the diversity of signs in the models we included in our simulations, this is not surprising. Based on our simulations, smaller policy moves would not much improve the case for coordination.

Another modification might be to coordinate only in those cases where the countries agree on the model they wish to use. Even assuming such agreement is possible, as noted above there is no reason why it should necessarily improve the incidence of gains from coordination, since agreeing on the model does not necessarily improve the chances that the chosen model is correct. However, for the subset of cases where the two countries do agree on a single model, the incidence of gains does happen to be somewhat higher for our simulations. The United States gains in 65 percent of the cases, while Europe gains in 59 percent of the cases.<sup>18</sup>

While our results indicate that coordination may frequently result in losses, we have said nothing of the magnitude of the losses or gains to cooperation. Specifically, it would be interesting to know whether there is an argument for cooperation based on the magnitude of the potential gains even in the best case when countries cooperate using the correct model. Oudiz and Sachs, 1984; Oudiz, 1985; Carlozzi and Taylor, 1985; Hughes Hallett, 1985; Canzoneri and Minford, 1986, and others who have estimated the gains from coordination have described them as small, even when positive as they must be in the conventional framework. But how small is "small"?

<sup>18</sup>Holtham and Hughes Hallett, 1987, p. 25, looking only at those cases where the two countries agree on the (perhaps wrong) model, confirm our finding: judged by the correct model, only slightly more than half the cases result in gains.

TABLE 6—GAINS TO UNILATERAL SWITCH TO TRUE MODEL FOR THE UNITED STATES UNDER NASH NONCOOPERATIVE SOLUTION

Model Subscribed to by the United States	Model Subscribed to by Europe			
	<i>MCM</i>	<i>VAR</i>	<i>OECD</i>	<i>LINK</i>
<i>MCM</i>				
Model Representing Reality				
<i>MCM</i> (0.0000)*	0.0000	0.0000	0.0000	0.0000
<i>VAR</i> (0.4323)	0.6069	0.0139	0.2412	0.3328
<i>OECD</i> (0.0128)	0.0011	-0.0000	0.0008	0.0007
<i>LINK</i> (0.0141)	0.0001	0.0008	0.0003	0.0001
<i>VAR</i>				
Model Representing Reality				
<i>MCM</i> (0.0000)	0.0964	0.0131	0.0920	0.0716
<i>VAR</i> (0.4323)	0.0000	0.0000	0.0000	0.0000
<i>OECD</i> (0.0128)	0.1019	0.0198	0.0946	0.0748
<i>LINK</i> (0.0141)	0.0447	0.0169	0.0389	0.0322
<i>OECD</i>				
Model Representing Reality				
<i>MCM</i> (0.0000)	0.0009	0.0000	0.0006	0.0006
<i>VAR</i> (0.4323)	0.4995	0.0137	0.2008	0.2750
<i>OECD</i> (0.0128)	0.0000	0.0000	0.0000	0.0000
<i>LINK</i> (0.0141)	0.0001	0.0009	-0.0000	0.0000
<i>LINK</i>				
Model Representing Reality				
<i>MCM</i> (0.0000)	0.0004	0.0001	0.0003	0.0003
<i>VAR</i> (0.4323)	0.5358	0.0179	0.2107	0.2899
<i>OECD</i> (0.0128)	0.0001	-0.0003	0.0001	0.0000
<i>LINK</i> (0.0141)	0.0000	0.0000	0.0000	0.0000

\*Gains to coordination to the United States assuming that all countries believe the same correct model.

In order to obtain a sense of how large or small the gain or loss to cooperation might be, we need a standard by which to judge these changes in welfare. We choose as a standard the gain to a single policymaker, who may previously have believed an incorrect model, of discovering the true model and unilaterally adjusting his policies accordingly while staying within the Nash non-cooperative equilibrium.

Table 6 shows the gains to the United States from a unilateral switch to the correct model by the United States. If the United States already has the correct model, the gains are zero. Otherwise, the gains are substantial. There is no *guaranteed* gain in utility to the United States when it switches to the correct model, as is illustrated by the occasional negative gains to the United States of a unilateral switch to the correct model. For example, if Europe were to believe the

OECD model, the United States would do better if it could play the OECD model as well even if it knew that the LINK model were correct. In these cases, the United States essentially loses bargaining power to Europe if it switches to the true model. But in most cases the gains from a unilateral switch are positive.

We show in parentheses in the left-hand column of the table numbers representing the gains to cooperation, under the assumption that all countries believe the same correct model to bias the case in favor of coordination. In the majority of cases, the gains to cooperation as shown are quite small compared to the gains to the unilateral switch to the correct model. As can be seen from Table 7, this is true of European gains when Europe makes the unilateral switch to the correct model as well. The gains from a unilateral switch to the correct model are

TABLE 7—GAINS TO UNILATERAL SWITCH TO TRUE MODEL FOR EUROPE UNDER NASH NONCOOPERATIVE SOLUTION

Model Subscribed to by the United States	Model Subscribed to by Europe			
	<i>MCM</i>	<i>VAR</i>	<i>OECD</i>	<i>LINK</i>
<i>MCM</i>				
Model Representing Reality				
<i>MCM</i> (0.0001) <sup>a</sup>	0.0000	1.3740	0.0665	0.0488
<i>VAR</i> (0.3216)	2.1378	0.0000	1.3004	1.4078
<i>OECD</i> (0.0078)	0.0193	0.2426	0.0000	0.0004
<i>LINK</i> (0.0038)	0.0141	0.2606	0.0004	0.0000
<i>VAR</i>				
Model Representing Reality				
<i>MCM</i> (0.0001)	0.0000	1.7716	0.0005	0.0181
<i>VAR</i> (0.3216)	0.8281	0.0000	0.8012	0.6667
<i>OECD</i> (0.0078)	0.0004	0.2706	0.0000	0.0007
<i>LINK</i> (0.0038)	0.0044	0.3293	0.0032	0.0000
<i>OECD</i>				
Model Representing Reality				
<i>MCM</i> (0.0001)	0.0000	1.3764	0.0572	0.453
<i>VAR</i> (0.3216)	1.9835	0.0000	1.2571	1.3287
<i>OECD</i> (0.0078)	0.0160	0.2424	0.0000	0.0002
<i>LINK</i> (0.0038)	0.0128	0.2609	0.0002	0.0000
<i>LINK</i>				
Model Representing Reality				
<i>MCM</i> (0.0001)	0.0000	1.3386	0.0595	0.0462
<i>VAR</i> (0.3216)	2.0375	0.0000	1.2690	1.3505
<i>OECD</i> (0.0078)	0.0171	0.2390	0.0000	0.0002
<i>LINK</i> (0.0038)	0.0133	0.2543	0.0002	0.0000

<sup>a</sup>Gains to coordination to the United States assuming that all countries believe the same correct model.

particularly large when a country erroneously believes the VAR model (or when the VAR is correct but the country erroneously believes any of the others).

### III. International Coordination of Monetary and Fiscal Policy Together

In this section we give each country a second tool, government expenditure— $g$  for the United States and  $g^*$  for Europe. We must add a third target variable for each country; otherwise each will be able to attain its optimal point regardless what the other country does. We choose the inflation rate. Now 24 multipliers are relevant from each model: the effects of  $m$ ,  $m^*$ ,  $g$ , and  $g^*$  on U.S. output, current account, and inflation and European output, current account, and inflation.

Table 8 reports the effects of fiscal expansion according to all 12 models. Table 2

reports the 24 multipliers for each of the four models. There is not as much disagreement regarding fiscal policy as monetary policy. A domestic fiscal expansion in most of the models is transmitted positively to the other country, via a domestic current account deficit. But a few models have fiscal or monetary expansion reducing the domestic price level rather than raising it.

We again assume that each country seeks to minimize a quadratic loss function. Rather than repeating our earlier points in algebraic form, we turn directly to the simulation results. As before, the weights and target optima are taken from Oudiz and Sachs (1984). The inflation target is zero for both the United States and Europe. Thus policy-makers will seek to reduce inflation, as well as to increase output and the current account.

Table 9 reports the Nash-bargaining solution. For one example, when the United States subscribes to the LIVPL model and

TABLE 8—FISCAL POLICY SIMULATION EFFECT IN SECOND YEAR OF INCREASE IN GOVERNMENT EXPENDITURE (1 PERCENT OF GNP)

	Y	CPI	<i>i</i>	Currency Value	CA	CA*	<i>i</i> *	CPI*	Y*
Fiscal Expansion in United States	United States				Rest of OECD				
	(in percent)	(in percent)	(Pts.)	(in percent)	(\$b)	(\$b)	(Pts.)	(in percent)	
MCM	+1.8	+0.4	+1.7	+2.8	-16.5	+8.9	+0.4	+0.4	+0.7
EEC <sup>a</sup>	+1.2	+0.6	+1.5	+0.6	-11.6	+6.6	+0.3	+0.2	+0.3
EPA <sup>b</sup>	+1.7	+0.9	+2.2	+1.9	-20.5	+9.3	+0.5	+0.3	+0.9
LINK	+1.2	+0.5	+0.2	-0.1	-6.4	+1.9	NA	-0.0	+0.1
LIVERPOOL	+0.6	+0.2	+0.4	+1.0	-7.0	+3.4	+0.1	+0.6	-0.0
MSG	+0.9	-0.1	+0.9	+3.2	-21.6	+22.7	+1.0	+0.5	+0.3
MINIMOD	+1.0	+0.3	+1.1	+1.0	-8.5	+5.5	+0.2	+0.1	+0.3
VAR <sup>c</sup>	+0.4	-0.9	+0.1	+1.2	-0.5	-0.2	-0.0	-0.0	-0.0
OECD	+1.1	+0.6	+1.7	+0.4	-14.2	+11.4	+0.7	+0.3	+0.4
TAYLOR <sup>c</sup>	+0.6	+0.5	+0.3	+4.0	NA	NA	+0.2	+0.4	+0.4
WHARTON	+1.4	+0.3	+1.1	-2.1	-15.4	+5.3	+0.6	-0.1	+0.2
DRI	+2.1	+0.4	+1.6	+3.2	-22.0	+0.8	+0.4	+0.3	+0.7

	Y	CPI	<i>i</i>	Currency Value	CA	CA*	<i>i</i> *	CPI*	Y*
Fiscal Expansion in Rest of OECD	Rest of OECD				United States				
	(in percent)	(in percent)	(Pts.)	(in percent)	(\$b)	(\$b)	(Pts.)	(in percent)	
MCM	+1.4	+0.3	+0.6	+0.3	-7.2	+7.9	+0.5	+0.2	+0.5
EEC <sup>a</sup>	+1.3	+0.8	+0.4	-0.6	-9.3	+3.0	+0.0	+0.1	+0.2
EPA <sup>b</sup>	+2.3	+0.7	+0.3	-0.7	-13.1	+4.7	+0.6	+0.3	+0.3
LINK	+1.2	+0.1	NA	-0.1	-6.1	+6.3	+0.0	+0.0	+0.2
LIVERPOOL	+0.3	+0.8	+0.0	+3.3	-17.2	+11.9	+0.8	+3.1	-0.5
MSG	+1.1	+0.1	+1.4	+2.9	-5.3	+10.5	+1.3	+0.6	+0.4
MINIMOD	+1.6	+0.2	+0.9	+0.6	-2.2	+3.2	+0.3	+0.2	+0.1
VAR <sup>c</sup>	+0.5	-0.3	-0.2	-2.4	+1.7	-2.6	+0.2	-0.1	+0.3
OECD	+1.5	+0.7	+1.9	+0.9	-6.9	+3.3	+0.3	+0.2	+0.1
TAYLOR <sup>c</sup>	+1.6	+1.2	+0.6	+2.7	NA	NA	+0.4	+0.9	+0.6
WHARTON	+3.2	-0.8	+0.8	-2.4	-5.5	+4.7	+0.1	-0.0	+0.0
DRI	NA	NA	NA	NA	NA	NA	NA	NA	NA

<sup>a</sup> Non-U.S. short-term interest rate NA; long-term rate reported instead.<sup>b</sup> Non-U.S. current account is Japan, Germany, United Kingdom, and Canada.<sup>c</sup> CPI NA. GNP deflator reported instead.

Europe to the EPA model, the resulting package of coordinated policy changes takes exactly the form urged by many economists in the 1980s: a U.S. fiscal contraction, accompanied by a fiscal expansion in the rest of the OECD and monetary expansion all around.<sup>19</sup> This package is considered desirable because it would depreciate the dollar and reduce the U.S. current account deficit (and European and Japanese surplus) with-

out causing a large world recession.<sup>20</sup> But most other possible kinds of policy packages occur as well, as can be seen in the table.<sup>21</sup>

<sup>20</sup> Table 8 in the NBER paper shows that according to the MSG model this change in the monetary/fiscal mix, though increasing non-U.S. output 0.1 percent and having the desired effect on the current accounts, would in fact reduce U.S. output 0.7 percent. There are several other combinations in the table where this same change in mix results from coordination, all of them involving the LIVPL model; but none of them shows quite the expected effects on the target variables.

<sup>21</sup> As in the case of coordination of monetary policy alone, there are a few cases of absurdly large changes, in particular the two combinations with the MSG and

<sup>19</sup> Examples include Olivier Blanchard and Rudiger Dornbusch, 1984; Layard et al., 1984; and Stephen Marris, 1985.

TABLE 9—THE COOPERATIVE BARGAIN (MONETARY AND FISCAL POLICIES)

Model Subscribed to by the United States	Model Subscribed to by Europe			
	<i>MCM</i>	<i>VAR</i>	<i>OECD</i>	<i>LINK</i>
<i>MCM</i>				
Bargaining Change in Policy				
Meur	-0.177	9.980	-55.755	0.869
Mus	1.259	4.010	-41.170	3.330
Geur	0.061	-2.892	0.052	-0.217
Gus	-0.157	-3.969	5.845	-0.955
Perceived Change in Targets				
Europe	<i>Y</i> -0.311	0.712	-11.830	-0.265
	<i>CA</i> -0.032	1.292	1.285	0.010
	<i>P</i> -0.260	-0.277	-1.365	-0.235
United States	<i>Y</i> 0.220	-7.048	-4.891	-0.579
	<i>CA</i> 0.053	1.015	-1.675	0.294
	<i>P</i> 0.084	-2.254	1.021	-0.136
Perceived Gain for				
Europe	0.0001	0.0326	0.0002	0.0059
United States	0.0002	0.0001	0.0001	0.0001
<i>VAR</i>				
Bargaining Change in Policy				
Meur	-258.885 <sup>a</sup>	43.693	207.547	-1.779
Mus	18.120	-17.723	-82.630	-7.732
Geur	39.164	-13.529	-27.962	-1.911
Gus	63.959	-11.088	-46.589	3.121
Perceived Change in Targets				
Europe	<i>Y</i> -0.652	-0.890	-25.266	-2.144
	<i>CA</i> 0.460	1.231	-10.996	0.339
	<i>P</i> -4.218	-1.846	-15.918	0.269
United States	<i>Y</i> -26.743	-8.678	-26.732	-5.658
	<i>CA</i> 13.579	-2.287	-13.287	-0.041
	<i>P</i> -14.363	2.014	0.143	-3.079
Perceived Gain for				
Europe	0.0001	0.0002	0.0000	0.0002
United States	0.0000	0.0001	0.0001	0.0003
<i>OECD</i>				
Bargaining Change in Policy				
Meur	213.120	47.154	-90.252	1.746
Mus	95.318	-20.088	22.653	11.040
Geur	-38.388	-20.510	11.924	1.035
Gus	-23.054	-4.444	-15.374	-3.387
Perceived Change in Targets				
Europe	<i>Y</i> -6.642	-4.012	-6.215	0.976
	<i>CA</i> 1.785	0.157	-5.653	-0.204
	<i>P</i> -3.068	-0.244	-4.800	-0.434
United States	<i>Y</i> 14.257	-13.796	-13.314	0.838
	<i>CA</i> 3.144	1.654	5.493	0.749
	<i>P</i> -10.157	-11.463	-3.019	0.063
Perceived Gain for				
Europe	0.0333	0.0000	0.0002	0.0001
United States	0.0001	0.0001	0.0000	0.0000
<i>LINK</i>				
Bargaining Change in Policy				
Meur	7.556	147.227	257.973	1.516
Mus	1.262	-53.522	-3.350	-1.796
Geur	-2.858	-55.997	-28.432	1.206
Gus	-0.219	-6.757	-3.605	-1.704
Perceived Change in Targets				
Europe	<i>Y</i> -1.542	-7.586	7.253	1.630
	<i>CA</i> 0.247	2.688	1.117	-0.263
	<i>P</i> -0.001	-2.942	-1.552	-0.057
United States	<i>Y</i> -0.331	-29.008	-4.400	-2.264
	<i>CA</i> -0.305	-2.619	1.920	0.586
	<i>P</i> -0.236	1.974	-1.467	-0.652
Perceived Gain for				
Europe	0.0003	0.0004	0.0001	0.0002
United States	0.0001	0.0001	0.0001	0.0001

<sup>a</sup>See fn. 17.



TABLE 10—TRUE GAINS FROM COORDINATION FOR THE UNITED STATES  
(MONETARY AND FISCAL POLICIES)

Model Subscribed to by the United States	Model Subscribed to by Europe			
	<i>MCM</i>	<i>VAR</i>	<i>OECD</i>	<i>LINK</i>
<i>MCM</i>				
Model Representing Reality				
<i>MCM</i>	0.0002	0.0001	0.0001	0.0001
<i>VAR</i>	-2.6394	10.6400	236.3649	-5.8930
<i>OECD</i>	-0.4994	-2.3167	14.8283	-0.0695
<i>LINK</i>	-0.5952	6.3941	25.8278	-1.7681
<i>VAR</i>				
Model Representing Reality				
<i>MCM</i>	-74.7859	198.0402	402.5238	13.1769
<i>VAR</i>	0.0000	0.0001	0.0001	0.0003
<i>OECD</i>	-25.1592	158.5107	373.6036	5.5733
<i>LINK</i>	-21.5520	39.1031	122.9517	10.4207
<i>OECD</i>				
Model Representing Reality				
<i>MCM</i>	-11.3838	29.9578	0.9014	0.0322
<i>VAR</i>	-268.2709	12.5137	347.2785	3.0793
<i>OECD</i>	0.0001	0.0001	0.0000	0.0000
<i>LINK</i>	-23.2350	-8.0046	-5.7921	-0.0567
<i>LINK</i>				
Model Representing Reality				
<i>MCM</i>	1.5302	233.1126	-125.2759	-1.2435
<i>VAR</i>	-6.7995	122.1154	-942.2249	8.2880
<i>OECD</i>	0.9647	119.0264	-43.4986	0.2318
<i>LINK</i>	0.0001	0.0001	0.0001	0.0001

TABLE 11—TRUE GAINS FROM COORDINATION FOR EUROPE  
(MONETARY AND FISCAL POLICIES)

Model Subscribed to by the United States	Model Subscribed to by Europe			
	<i>MCM</i>	<i>VAR</i>	<i>OECD</i>	<i>LINK</i>
<i>MCM</i>				
Model Representing Reality				
<i>MCM</i>	0.0001	-3.3306	-5.6736	-3.1598
<i>VAR</i>	-2.5132	0.0326	333.6467	-8.2613
<i>OECD</i>	0.1306	13.3216	0.0002	0.7635
<i>LINK</i>	0.0093	43.8440	44.4689	0.0059
<i>VAR</i>				
Model Representing Reality				
<i>MCM</i>	0.0001	6.2655	270.6773	-4.8492
<i>VAR</i>	217.4242	0.0002	209.1031	25.7900
<i>OECD</i>	-367.2546	327.4278	0.0000	1.2868
<i>LINK</i>	-973.6363	327.6573	980.1376	0.0002
<i>OECD</i>				
Model Representing Reality				
<i>MCM</i>	0.0333	17.3204	176.5559	12.6783
<i>VAR</i>	-549.5166	0.0000	241.4204	-29.8594
<i>OECD</i>	142.8152	199.9979	0.0002	1.4537
<i>LINK</i>	289.0842	313.7216	119.2115	0.0001
<i>LINK</i>				
Model Representing Reality				
<i>MCM</i>	0.0003	141.2746	-696.8871	-0.7884
<i>VAR</i>	-7.2841	0.004	-1276.6437	-10.8551
<i>OECD</i>	16.4437	828.7074	0.0001	-1.7457
<i>LINK</i>	16.7499	1143.2227	-1016.8853	0.0002

Tables 10 and 11 show the true gains from coordination for the United States and Europe, respectively. Again we find that coordination necessarily improves U.S. welfare if the U.S. model turns out to be the correct one, and European welfare if the European model turns out to be the correct one, but that otherwise welfare can go down. Of the total 1,000 combinations of all ten models, the United States has gains in 494 cases, losses in 398, and no perceptible effect in 108. Europe has gains in 477 cases, losses in 418, and no effect in 105. If we take only the 810 combinations where neither country is correct, bargaining results in U.S. gains in 432 cases and losses in 357, and for Europe gains in 408 cases and losses in 376. Thus the odds for successful coordination appear to be no better when policymakers can take advantage of the monetary-fiscal mix than when the degree of monetary ease is alone at stake.

#### IV. Extensions with Uncertainty

So far we have made the simplest assumptions to examine the topic at hand. Some readers of earlier versions have suggested that, in a world in which different models abound, it is not sensible to assume that each policymaker acts as if he knows with certainty to which model his opponent subscribes or even which model he himself considers to be correct. We now briefly consider extensions in each of these two directions in turn.

To begin with, we retain the assumption that each policymaker believes in his own model with certainty, but we allow for uncertainty regarding the other's model. A reason for such uncertainty regarding the other's model might be that several models might be believed by different policymakers within the other country's government. The model which will actually be used in setting policy is the unknown outcome of a political pro-

cess within the other country. The policymaker will set his policies so as to maximize expected welfare, a weighted average of the economic consequences of each of the policy settings that the foreign government would choose under each of the possible models to which it might subscribe. The foreign government's policy settings in turn will depend, not just on its model, but also on its beliefs about what the first country's model, and therefore its actions, might be.

The U.S. central bank chooses  $m_j$  to minimize

$$\sum_{i=1}^{10} \pi_{ij}^* W_i(m_j, m_i^*),$$

where  $\pi_{ij}^*$  is the U.S. estimate of the probability that Europe believes in model  $i$  given that the United States believes model  $j$  and  $m_i^*$  is the money supply Europe will pick if it believes in model  $i$ . If the U.S. central bank believes in, for example, model 1, then the first-order condition is similar to equation (7), but with the foreign money supply replaced by a weighted average of the possibilities.

$$(7') \quad m_1 = M_1 + N_1 \sum_{i=1}^{10} \pi_{i1}^* m_i^*$$

or

$$m_1 = M_1 + N_1 (\pi_1^{*'} m^*),$$

where  $\pi_1^{*'}$  is the row vector  $\pi_{i1}^*$  and  $m^*$  is the column vector of  $m_i^*$  (each for  $i=1, 10$ , assuming ten possible models).

Similarly the European central bank chooses  $m_k^*$  to minimize

$$\sum_{i=1}^{10} \pi_{ik} W_i^*(m_i, m_k^*),$$

where  $\pi_{ik}$  is the European estimate of the probability that the United States believes in model  $i$  given that Europe believes model

MCM models. The explanation, again, is that these changes offset absurdly large changes implied by the move from the baseline to the Nash equilibrium.

$k$ ,<sup>22</sup> and  $m_i$  is the money supply the United States will pick if it believes in model  $i$ . If the European central bank believes in, for example, model 2, then the first-order condition is

$$(8') \quad m_2^* = Q_2 + R_2(\pi_k' m),$$

where  $\pi_k'$  is the row vector of  $\pi_{ik}$  and  $m$  is the column vector of  $m_i$ . We have one version of equation (7') for each of the ten models in which the U.S. central bank might believe, giving

$$(7'') \quad m = M + N(\pi^* m^*)$$

and similarly for Europe,

$$(8'') \quad m^* = \underline{Q} + \underline{R}(\pi m),$$

where  $\pi^*$  is the  $(10 \times 10)$  matrix of the  $\pi_j^*$ ,  $\underline{M}$  is the  $(10 \times 10)$  matrix of the  $\underline{m}_j$ ,  $\underline{M}$  and  $\underline{Q}$  are the  $(10 \times 1)$  vector forms of  $M_i$  and  $Q_i$ , and  $\underline{N}$  and  $\underline{R}$  are  $(10 \times 10)$  diagonal matrices with the  $N_i$  and  $R_i$  on the diagonal. Substituting and solving,

$$(9') \quad m = [I - N\pi^*R\pi]^{-1}[\underline{M} + N\pi^*Q],$$

$$(10') \quad m^* = [I - R\pi N\pi^*]^{-1}[\underline{Q} + R\pi M],$$

where  $I$  is the  $(10 \times 10)$  identity matrix.

Equations (9')–(10') represent the  $10 \times 10$  computable noncooperative solutions for the  $10 \times 10$  combinations of models in which the two policymakers could believe. As a concrete example we try putting equal weight on each of our ten Brookings' models:  $\pi_i = \pi_i^* = 1/10$  ( $i = 1, 10$ ). The bargaining solution remains the same as before, assuming that

an enforcement mechanism is designed such that each policymaker must reveal his model as part of the cooperative bargain. As before we calculate in each case the gain or loss in welfare entailed in the move from one equilibrium to the other, where the true effect of any given pair of money supplies is judged by each of the ten models in turn.

The simulations indicate that the noncooperative point when each country believes its own model with certainty but averages over the possible models followed by the foreign country, is quite similar to the noncooperative point under certainty.<sup>23</sup> As in the earlier sections, the interesting question, under the assumption that each player averages to estimate the other's model, is the effect of coordination. Coordination under averaging improves U.S. welfare in 600 cases out of the total 1,000 combinations, against 398 losses and 2 cases with no significant change in welfare. For Europe, welfare improves in 643 cases, falls in 355 cases and has no significant change in 2 cases.

The second extension relaxes the assumption that each policymaker acts as if he were certain as to the correct model. We assume rather that policymakers assign weight to the possibility that each of the ten models may be true, and choose their policies so as to maximize expected welfare. To preserve some disagreement about models, we could assume that each puts primary weight on a favorite model of his own, but also puts some weight on the other models (perhaps with larger weight on the favorite model of the other player, on the theory that he must have access to some independent information). Here we consider, instead, the simple case of uniform weights. As a result, each will be playing by the same "compromise" model. If heavier weight were placed on particular models, the solution would lie between the results shown here and solution under certainty.

The noncooperative solution under averaging by both policymakers over the possible

<sup>22</sup>The probabilities,  $\pi_j^*$  and  $\pi_k$ , are conditional probabilities, given the beliefs of the United States ( $j$ ) or Europe ( $k$ ). These conditional probabilities are formed using Bayes rule from some underlying probability distribution over models which is known by both countries' policymakers.

<sup>23</sup>The tables showing the results for a subset of the models are reported in Frankel (1988). They are omitted here to save space.

correct models lies farther from the noncooperative solution under certainty (as in Section II) than does the noncooperative solution when each policymaker averages over the other's model (as in the first extension). In this case, all the multipliers change, not just the foreign multipliers, so a larger move would be expected.

As before, the main interest is in characterizing the move to the cooperative point. But two types of cooperation are possible. In the first case, the cooperative point is the Nash-bargaining solution given that each country believes the "compromise" model. If this type of cooperation is compared to the noncooperative averaging solution, there is a true gain for Europe in more than half the cases, but a loss for the United States in the majority of cases: the United States gains in 200 cases, but loses in 800 cases, while Europe gains in 600 cases and loses in 400 cases.<sup>24</sup>

The second type of coordination would suppose that the two countries stubbornly continue to believe their own favorite models, but for the sake of compromise, the two countries agree simply to average over the possible correct models and play either noncooperatively or cooperatively. If this concept of cooperation is used, both countries gain in more cases than they lose. If the noncooperative averaging point is compared to the noncooperative point under certainty, the move results in gains for the United States in 568 cases and losses in 432 cases, while the move results in gains for Europe in 513 cases and losses in 487 cases. The probable reason that averaging usually raises welfare is the statistical principle that the average of ten numbers is closer to the individual numbers, on average, than the individual numbers are to each other. The principle does not apply directly, because each policymaker's having a better estimate of the "true" parameters does not necessarily imply that the noncooperative equilibrium will be better, but it seems to work here.

These extensions are more elaborate models of the Nash noncooperative equilibrium.

None offers an evident reason for altering our conclusion that the bargaining solution is as likely to reduce welfare as to improve it. But more definitions of cooperation should be investigated, including exchange of information over time to allow learning regarding the correct model. The scope for useful international cooperation remains wide, provided it is defined more broadly than in the conventional bargaining sense explored in the first sections of this paper.

## REFERENCES

- Bergsten, C. Fred et al., *Promoting World Recovery: A Statement on Global Economic Strategy by Twenty-Six Economists from Fourteen Countries*, Institute for International Economics, December 1982.
- Blanchard, Olivier and Dornbusch, Rudiger, "U.S. Deficits, the Dollar and Europe," *Banca Nazionale del Lavoro Quarterly Review*, March 1984, 148, 89-113.
- Brainard, William, "Uncertainty and the Effectiveness of Policy," *American Economic Review*, May 1967, 57, 411-25.
- Branson, William, "The Limits of Monetary Coordination as Exchange Rate Policy," *Brookings Papers on Economic Activity*, 1: 1986, 175-88.
- Bryant, Ralph, Henderson, Dale, Holtham, Gerald, Hooper, Peter and Symansky, Steven, eds., *Empirical Macroeconomics for Interdependent Economics*, Washington: The Brookings Institution, 1988.
- Buiter, Willem and Marston, Richard, *International Economic Policy Coordination*, New York: Cambridge University Press, 1985.
- Canzoneri, Matthew and Gray, JoAnna, "Monetary Policy Games and the Consequences of Non-Cooperative Behavior," *International Economic Review*, October 1985, 26, 547-64.
- and Minford, Patrick, "When Policy Coordination Matters: An Empirical Analysis," Center for Economic Policy Research Discussion Paper No. 119, July 1986.
- Carlozzi, Nicholas and Taylor, John, "International Capital Mobility and the Coordination of Monetary Rules," in *Exchange*

<sup>24</sup>Again the tables are reported in Frankel (1988).

- Rate Management under Uncertainty*, J. Bhandari, ed., Cambridge: MIT Press, 1985.
- Cooper, Richard, "Economic Interdependence and Coordination of Economic Policies," in *Handbook in International Economics*, Vol. II, R. Jones and P. Kenen, eds, Amsterdam: North-Holland, 1985.
- \_\_\_\_\_, "International Cooperation in Public Health as a Prologue to Macroeconomic Cooperation," *Brookings Discussion Papers in International Economics*, 44, Washington: The Brookings Institution, March 1986.
- Feldstein, Martin, "The World Economy," *The Economist*, June 1983.
- Fischer, Stanley, "International Macroeconomic Policy Coordination," NBER Working Paper No. 2244, May 1987, in *International Policy Coordination*, M. Feldstein, ed., Chicago: University of Chicago Press (forthcoming 1988).
- Frankel, Jeffrey, (1986a), "The Sources of Disagreement Among International Macro Models and Implications for Policy Coordination," NBER Working Paper No. 1925, May 1986, revised and forthcoming in *Empirical Macroeconomics for Interdependent Economics*, Ralph Bryant et al., eds., as "Ambiguous Macroeconomic Policy Multipliers, in Theory and in Twelve Econometric Models," and (1986b), "The Implications of Conflicting Models for Coordination Between Monetary and Fiscal Policy-Makers."
- \_\_\_\_\_, "Obstacles to International Macroeconomic Policy Coordination," International Monetary Fund Working Paper 87/28, *Studies in International Finance*, Princeton University, forthcoming 1988.
- Ghosh, Atish, "International Policy Coordination in an Uncertain World," *Economic Letters*, 1986, 21, 271-6.
- Ghosh, Swati and Ghosh, Atish, "International Policy Coordination When the Model is Unknown," mimeo., Geneva, Switzerland, December 1986.
- Hamada, Koichi, "A Strategic Analysis of Monetary Interdependence," *Journal of Political Economy*, August 1976, 84, 77-99.
- Holtham, Gerald, "International Policy Coordination: How Much Consensus Is There?," *Brookings Discussion Papers in International Economics*, 50, Washington: The Brookings Institution, September 1986.
- \_\_\_\_\_, and Hughes Hallet, A. J., "International Policy Cooperation and Model Uncertainty," Center for Economic Policy Research Discussion Paper No. 190, London, July 1987. "How Much Could International Coordination of Economic Policies Achieve? An Example from U.S.-E.E.C. Policy-Making," Center for Economic Policy Research Discussion Paper No. 77, London, 1985.
- Ishii, Naoko, McKibbin, Warwick and Sachs, Jeffrey, "The Economic Policy Mix, Policy Cooperation, and Protectionism: Some Aspects of Macroeconomic Interdependence Among the United States, Japan and Other OECD Countries," *Journal of Policy Modeling*, 1985, 7, 533-72.
- Kehoe, Patrick, "International Policy Cooperation May Be Undesirable," Federal Reserve Bank of Minneapolis Research Department Staff Report No. 103, February 1986.
- Layard, Richard et al., "The Case for Unsustainable Growth," *CEPS Discussion Papers*, Brussels: Center for European Policy, February 1986.
- Marris, Stephen, *Deficits and the Dollar: The World Economy at Risk*, Institute for International Economics, Policy Analyses in International Economics No. 14, December 1985.
- Miller, Marcus and Salmon, Mark, "Dynamic Games and the Time Inconsistency of Optimal Policy in Open Economies," *Economic Journal*, 1985, 85, Suppl., 124-37.
- Mussa, Michael, "Macroeconomic Interdependence and the Exchange Rate Regime," in *International Economic Policy: Theory and Evidence*, R. Dornbusch and J. Frankel, eds., Baltimore: Johns Hopkins, 1979.
- Oudiz, Gilles, "European Policy Coordination: An Evaluation," Centre for Economic Policy Research Discussion Paper No. 81, October 1985.
- \_\_\_\_\_, and Sachs, Jeffrey, "Macroeconomic Policy Coordination Among the Industrialized Economies," *Brookings Papers on*

- Economic Activity*, 1: 1984, 1-64.
- Rogoff, Kenneth**, "Can International Monetary Policy Cooperation be Counterproductive?," *Journal of International Economics*, February 1985, 18, 199-217.
- Roubini, Nouriel**, "International Policy Coordination and Model Uncertainty," unpublished paper, Harvard University, 1986.
- Svensson, Lars and van Wijnbergen, Sweder**, "International Transmission of Monetary Policy," NBER Summer Institute Paper, August 1986.
- Taylor, John**, "International Coordination in the Design of Macroeconomic Policy Rules," *European Economic Review*, June/July 1985, 28, 53-81.

# Dynamic Strategic Monetary Policies and Coordination in Interdependent Economies

By STEPHEN J. TURNOVSKY, TAMER BASAR, AND VASCO D'OREY\*

*This paper develops dynamic strategic monetary policies using a standard two-country macro model under flexible exchange rates. The equilibria considered include feedback Nash and feedback Stackelberg, both of which are compared to the Pareto-optimal cooperative equilibrium. The optimal policies are obtained as feedback rules in which real money supplies are adjusted to movements in the real exchange rate. The properties of these policies and their welfare implications are analyzed using numerical simulations.*

With the increasing interdependence between national economies, there has been a growing interest in problems of strategic policymaking and international policy coordination. Research into these issues began with the seminal work of Koichi Hamada (1976), who analyzed issues of monetary policy under Cournot and Stackelberg behavior. His approach was a static one and was based on a fixed exchange rate. His contribution has recently been extended by various authors including Michael Jones, 1983; Matthew Canzoneri and Jo Anna Gray, 1985; Stephen Turnovsky and Vasco d'Orey, 1986.

In this paper we consider the problem of strategic monetary policymaking within a dynamic framework. The basic model we employ is a two-country version of the standard Rudiger Dornbusch (1976) model in which the policymakers in the two economies seek to optimize their respective objec-

tive functions, taken to be intertemporal quadratic cost functions defined in terms of deviations in output from its natural rate level, on the one hand, and the rate of inflation of the domestic consumer price index (CPI) on the other.

The consideration of these issues within a dynamic context is obviously important. Strategic policies, which are optimal from a short-run viewpoint, may, however, generate intertemporal tradeoffs which over time prove to be adverse. In fact, our results below will suggest this to be the case. Furthermore, the extension to a dynamic framework emphasizes new issues such as the information structure and the corresponding equilibrium concepts. The equilibria we consider are all feedback solutions, in which the policies at each stage make use of current information on key economic variables such as prices and exchange rates, which under our assumptions are observable at that time. Using such information we analyze and compare two noncooperative equilibria, which we consider to be of interest: (i) feedback Nash, and (ii) feedback Stackelberg.<sup>1</sup>

\*University of Washington, Seattle, WA 98195, and National Bureau of Economic Research, Cambridge, MA 02138; University of Illinois at Urbana-Champaign; and Universidade Nova de Lisboa, Lisbon, Portugal, respectively. Previous versions of this paper were presented at the SEDC Conference held at Imperial College, London, June 1985, the Summer Workshop, held at the University of Warwick, Coventry, 1985, and the ASSA meetings in New York, December 1985. The comments of a referee are gratefully acknowledged. This research was supported in part by grant no. SES-8409886 from the National Science Foundation.

<sup>1</sup>In an expanded version of this paper, the feedback consistent conjectural variations (CCV) equilibrium is also considered; see Chaim Fershtman and Morton Kamien, 1985; Tamer Basar, 1985. This is a new equilibrium concept in dynamic game theory and is a generalization of the static CCV equilibrium concept intro-

A basic question throughout the recent policy discussion concerns the gains from policy coordination. We address this issue by deriving the Pareto-optimal cooperative equilibrium, where the two policymakers agree to minimize their aggregate joint welfare costs. This equilibrium is then compared with the two noncooperative equilibria.

Any strategic policy problem must be generated by some disturbance to an initial equilibrium situation, thereby creating a conflict for the two policymakers. In the present analysis, this is taken to be an initial misalignment in the real exchange rate. In general, this may be the result of a variety of underlying causes. Here, it most naturally reflects past differences in monetary policy, resulting in differential price movements in the two economies and leading to the inherited exchange rate misalignment. The policy problem is therefore to return to equilibrium with a minimum of welfare losses.<sup>2</sup>

The analysis is based on two symmetric economies. This has the advantage of simplifying the feedback rules, with the real money supply in each economy being adjusted to the real exchange rate. Our procedure is to derive analytical expressions for the optimal policies. We then use these analytical expressions to compute values for the policy rules and the welfare gains, using the numerical estimates of the parameters of the model.

This is not the first study to apply dynamic game theory to problems of international macroeconomic policymaking. Indeed, the area has recently begun to receive increased attention and recent work by Marcus Miller and Mark Salmon, 1985; David Currie

and Paul Levine, 1985; Gilles Oudiz and Jeffrey Sachs, 1985; John Taylor, 1985; and Andrew Hughes Hallett, 1984, in particular, should be noted. These contributions can be generally characterized as being variants of the standard Keynesian IS-LM Phillips curve framework, and for reasons of analytical complexity employ numerical simulation methods. While this characterization is also true of the present study, it also differs in many key respects.<sup>3</sup> One of these is in the types of strategic equilibria considered. As noted, we focus on feedback solutions, which are determined using dynamic programming methods and are known to be time consistent. By contrast, authors such as Miller and Salmon, Oudiz and Sachs, and Hughes Hallett emphasize the contrast between time-consistent and time-inconsistent solutions.

Much of the literature focuses on the gains from cooperation. In this regard, Miller and Salmon present an example in which cooperation may actually lead to welfare losses, a finding also obtained previously by K. Rogoff (1985), although for substantially different reasons. By contrast, this study, like Oudiz and Sachs, 1985, and Taylor, 1985, finds cooperation to yield welfare gains. These are found to be of the order of around 6 to 10 percent, which are similar in magnitude to those obtained by Taylor, but larger than those suggested by Oudiz and Sachs. Also, in contrast to these latter authors, who find the cooperative solution to be more inflationary than the noncooperative, we find just the reverse in fact to be the case. This result would appear to be generally consistent with Taylor whose multicountry analysis does not yield a uniform pattern in this respect. Fi-

duced by Timothy Bresnahan, 1981; Martin Perry, 1982; and Kamien and Nancy Schwartz, 1983. Andrew Hughes Hallett, 1984, considers arbitrary, but not consistent, conjectural variations in a dynamic policy game framework. Unfortunately, space limitations preclude a detailed discussion of this equilibrium. But some of our results are noted in footnotes at appropriate places.

<sup>2</sup>Viewed in this way, the problem may be regarded as being a strategic analogue to the problem of the optimal reduction of inflation originally considered by E. S. Phelps (1967) and studied by several authors since.

<sup>3</sup>While these papers belong to the same generic class, they differ in terms of their technical details. For example, Miller and Salmon use continuous time, with the interest rate being the policy variable. Oudiz and Sachs introduce more sluggish wage behavior, the result of which is that the optimal monetary rule depends upon a greater set of lagged variables. Currie and Levine (1985) have a stochastic model, but consider a set of simple, but not fully optimal, monetary feedback rules, while Taylor assumes staggered wages and prices. Finally, several of the authors consider open-loop, as well as feedback rules.



nally, our approach differs from the previous literature in two further aspects. First, to avoid the danger of excessive reliance on specific parameter values, a much more detailed sensitivity analysis is conducted, the result of which is to suggest that our findings are in fact quite robust across parameter sets. Second, unlike previous authors, our analysis stresses the contrast between the results obtained in the present dynamic analysis with those obtained previously for the more familiar short-run (one-period) model. The differences are shown to be quite striking, highlighting the intertemporal, as well as the intratemporal, tradeoffs involved.

### I. The Theoretical Framework

The analysis of this paper is based on the following two-country macroeconomic model, which is a direct extension of the Dornbusch (1976) framework. It describes two identical countries, each specializing in the production of a distinct good and trading a single common bond. It assumes perfect foresight and is expressed, using discrete time, by the following set of equations

$$(1) \quad Y_t = d_1 Y_t^* - d_2 [I_t - (P_{t+1} - P_t)] \\ + d_3 (P_t^* + E_t - P_t) \\ 0 < d_1 < 1, d_2 > 0, d_3 > 0$$

$$(1') \quad Y_t^* = d_1 Y_t - d_2 [I_t^* - (P_{t+1}^* - P_t^*)] \\ - d_3 (P_t^* + E_t - P_t)$$

$$(2) \quad M_t - P_t = e_1 Y_t - e_2 I_t \\ e_1 > 0, e_2 > 0$$

$$(2') \quad M_t^* - P_t^* = e_1 Y_t^* - e_2 I_t^*$$

$$(3) \quad I_t = I_t^* + E_{t+1} - E_t$$

$$(4) \quad C_t = \delta P_t + (1 - \delta)(P_t^* + E_t) \\ 1 > \delta > 1/2$$

$$(4') \quad C_t^* = \delta P_t^* + (1 - \delta)(P_t^* - E_t)$$

$$(5) \quad P_{t+1} - P_t = \gamma Y_t \\ \gamma > 0$$

$$(5') \quad P_{t+1}^* - P_t^* = \gamma Y_t^*,$$

where  $Y$  = real output, in logarithms, measured as a deviation about its natural rate level;

$P$  = price of domestic output, expressed in logarithms;

$C$  = consumer price index, expressed in logarithms;

$E$  = exchange rate (measured in terms of units of foreign currency per unit of domestic currency), measured in logarithms;

$I$  = nominal interest rate, measured in natural units;

$M$  = nominal money supply, expressed in logarithms.

Domestic variables are unstarred; foreign variables are denoted with asterisks. We shall also refer to these as Country 1 and Country 2, respectively.

Equations (1) and (1') describe equilibrium in the two-goods markets. Output depends upon the real interest rate, output in the other country, and the relative price. The corresponding effects across the two economies are identical, with relative price influencing demand in exactly offsetting ways. The money market equilibrium conditions in the two economies are standard and are described by (2) and (2'), respectively.<sup>4</sup> The perfect substitutability between domestic and foreign bonds is described by the interest rate parity condition (3). Equations (4) and (4') describe the consumer price index (CPI) in the two economies. They embody the assumption that the proportion of consumption  $\delta$  spent on the respective home good is the same in the two economies.<sup>5</sup> Note that the real interest rate in (1) and (1') and the real money supplies in (2) and (2') are deflated by the output price of their respective economies. Little would be

<sup>4</sup>We maintain the usual assumption that residents of one country do not hold the currency of the other country.

<sup>5</sup>We assume  $1 > \delta > \frac{1}{2}$ , so that residents in both countries have a preference for their own good. Note that the real interest rate in (1) and (1') and the real money supplies in (2) and (2') are deflated by the output price of their respective economies. Little would be changed, except for additional detail, if the deflators were in terms of their respective CPI's.

changed, except for additional detail, if the deflators were in terms of their respective CPI's. Equations (5) and (5') define the price adjustment in the two economies in terms of Phillips curve relationships, with prices responding with a one-period lag to demand. On the other hand, the assumption of perfect foresight is embodied in the future price level and future exchange rate appearing in the real interest rate in (1), (1'), and the interest rate parity relationship (3).

Equations (1) to (5) describe the structure of the two economies. The policymakers in these economies are assumed to have intertemporal objective functions

$$(6) \quad \sum_{t=1}^T [aY_t^2 + (1-a)(C_{t+1} - C_t)^2] \rho^{t-1}$$

$$0 < a < 1 \quad 0 < \rho < 1$$

$$(6') \quad \sum_{t=1}^T [aY_t^{*2} + (1-a)(C_{t+1}^* - C_t^*)^2] \rho^{t-1}$$

which they seek to optimize. That is, each policymaker chooses to minimize an intertemporal cost function. The cost incurred at each point of time is quadratic, defined in terms of deviations in output from its equilibrium, natural rate, level, and the rate of inflation of the domestic cost of living. The relative weights attached to these components of the objective functions are  $a$  and  $1-a$ , respectively. Total cost to be minimized is a discounted sum of the costs incurred at each period, with  $\rho$  denoting the discount rate. Equations (1) to (5) may be solved for  $Y_t$ ,  $Y_t^*$ , and  $E_{t+1} - E_t$ , as follows

$$(7a) \quad Y_t = \phi_1 m_t + \phi_2 m_t^* + \phi_3 s_t$$

$$(7b) \quad Y_t^* = \phi_2 m_t + \phi_1 m_t^* - \phi_3 s_t$$

$$(7c) \quad E_{t+1} - E_t = -\beta_1 m_t + \beta_1 m_t^* + \beta_3 s_t,$$

where  $s_t \equiv P_t^* + E_t - P_t$  denotes the relative price (real exchange rate) at time  $t$ ;  $m_t \equiv M_t - P_t$ ,  $m_t^* \equiv M_t^* - P_t^*$  denote the real stocks

of money at home and abroad, at time  $t$ ,

$$\phi_1 \equiv \frac{d_2}{2} \left[ \frac{1}{D} + \frac{1}{D'} \right];$$

$$\phi_2 \equiv \frac{d_2}{2} \left[ \frac{1}{D} - \frac{1}{D'} \right];$$

$$\phi_3 \equiv \frac{e_2 d_3}{D'}.$$

$$\beta_1 \equiv \frac{1 + d_1 - d_2 \gamma}{D'};$$

$$\beta_2 \equiv \frac{2e_1 d_3}{D'}.$$

$$D \equiv e_2(1 - d_1 - d_2 \gamma) + e_1 d_2;$$

$$D' \equiv e_2(1 + d_1 - d_2 \gamma) + e_1 d_2.$$

We assume that  $1 - d_1 - d_2 \gamma > 0$ , implying that the IS curve of the aggregate world economy is downward sloping. It follows that

$$D' > D > 0,$$

and hence

$$\phi_1 > \phi_2 > 0.$$

Taking the differences of the cost of living equations (4) at two consecutive points in time, and using (5), (5'), and (7a)–(7c), the rates of inflation of the CPI become

$$(8a) \quad C_{t+1} - C_t = \eta_1 m_t + \eta_2 m_t^* + \eta_3 s_t$$

$$(8b) \quad C_{t+1}^* - C_t^* = \eta_2 m_t + \eta_1 m_t^* - \eta_3 s_t,$$

where

$$\eta_1 \equiv \gamma [\delta \phi_1 + (1 - \delta) \phi_2] - \beta_1 (1 - \delta),$$

$$\eta_2 \equiv \gamma [\delta \phi_2 + (1 - \delta) \phi_1] + \beta_1 (1 - \delta),$$

$$\eta_3 \equiv \gamma \phi_3 (2\delta - 1) + \beta_2 (1 - \delta).$$

The optimal policy problem confronting each of the policymakers is to choose their

respective money supplies to minimize their cost functions (6) and (6') subject to constraints (8a), (8b), (9a), (9b). Given the assumption that prices move gradually at home and abroad, we assume that both  $P_t$  and  $P_t^*$  are observed at time  $t$ . Thus it is convenient to treat the monetary control variables as being the real quantities  $m, m^*$ . Second, we assume that the current nominal exchange rate  $E_t$  is observed instantaneously and can therefore be monitored by the monetary authorities.<sup>6</sup> Thus the relative price,  $s_t$ , is observable to both policymakers at time  $t$ , and in fact the optimal monetary policies will be obtained as feedback solutions in terms of  $s_t$ . Combining equations (5), (5'), and (7c),  $s_t$  follows the path

$$(9) \quad s_{t+1} = cs_t + bm_t - bm_t^*$$

where

$$c \equiv 1 + \beta_2 - 2\gamma\phi_3$$

$$b \equiv -(1 + d_1)/D'.$$

In considering equation (9), it should be noted that  $m_t, m_t^*$  denote *real* money stocks, which given that the price levels  $P_t, P_t^*$  are constrained to move sluggishly, can be treated as policy variables. With forward-looking variables, such as the nominal exchange rate, one normally expects the dynamics of such a system to involve a saddlepoint. This is in fact also the case here, when one makes the usual assumption that the *nominal* money supplies remain fixed, or follow some exogenous path. In this case to specify the dynamics, one needs to combine (9) with the price adjustment equations (5), (5') together with (7a), (7b). The result of

this is the following matrix equation system

$$(10) \quad \begin{bmatrix} s_{t+1} \\ P_{t+1} \\ P_{t+1}^* \end{bmatrix} = \begin{bmatrix} c & -b & b \\ \phi_3 & 1 - \gamma\phi_1 & -\gamma\phi_2 \\ -\phi_3 & -\gamma\phi_2 & 1 - \gamma\phi_1 \end{bmatrix} \begin{bmatrix} s_t \\ P_t \\ P_t^* \end{bmatrix} + \begin{bmatrix} b & -b \\ \gamma\phi_1 & \gamma\phi_2 \\ \gamma\phi_2 & \gamma\phi_1 \end{bmatrix} \begin{bmatrix} M_t \\ M_t^* \end{bmatrix}.$$

Under plausible conditions, for given  $M_t, M_t^*$ , this will have two stable and one unstable root, with the real exchange rate jumping (via the nominal rate) to ensure that the system is always following a stable path.<sup>7</sup> But with endogenous feedback policy, stability can be accomplished through appropriate adjustments in the policy variables. The unstable root may be eliminated from the system, ruling out the need for the exchange rate to undergo endogenous jumps.

In order to see how the strategic problem is generated, suppose that prior to time 1 the two monetary authorities have been allowing their respective nominal money stocks to follow exogenous time paths. From equations (10) we see that the real exchange rate  $s_{t+1}$  is generated by

$$\begin{aligned} s_{t+1} &= cs_t - b[P_t - P_t^*] + b[M_t - M_t^*], \\ &= cs_t + b[(M_t - P_t) - (M_t^* - P_t^*)], \\ &= [1 - \gamma(\phi_1 - \phi_2) + c]s_t \\ &\quad + [c[1 - \gamma(\phi_1 - \phi_2)] + 2b\phi_3]s_{t-1} \\ &\quad + b[\Delta M_t - \Delta M_t^*]. \end{aligned}$$

Assume that at some distant time in the past, the world economy was in long-run

<sup>6</sup>This assumption of the instantaneous observability of the exchange rate is the standard one in the current exchange market intervention literature.

<sup>7</sup>For the analysis of exogenous policy shocks in such a model, see S. J. Turnovsky (1986).

equilibrium with  $s=0$ . It is then evident from this equation (or more precisely its stable solution) that the misalignment in the real exchange rate, which forms the starting point for the present strategic analysis, reflects differential monetary policies in the two economies over the entire prior period and in particular how these manifest themselves in differential real money stocks. These disturbances can be either transitory, lasting just one period; or they can be sustained differences in monetary growth rates. These will simply result in different values of the real exchange rate at time 1. Even though any stable adjustment path will ensure the ultimate reattainment of the equilibrium exchange rate, the introduction of strategic behavior at some arbitrary point can be viewed as an attempt to accelerate this adjustment process. In a more general model, a misaligned initial exchange rate could also reflect other factors, such as differential fiscal policies or supply shocks to the two economies.

The dynamic optimization problem faced by the two policymakers may be summarized as

$$(11) \quad \text{Min } J_T$$

$$= \sum_{t=1}^T \left[ aY_t^2 + (1-a)(C_{t+1} - C_t)^2 \right] \rho^{t-1}$$

$$(12) \quad \text{subject to } Y_t = \phi_1 m_t + \phi_2 m_t^* + \phi_3 s_t$$

$$(13) \quad C_{t+1} - C_t = \eta_1 m_t + \eta_2 m_t^* + \eta_3 s_t$$

and

$$(11') \quad \text{Min } J_T^*$$

$$= \sum_{t=1}^T \left[ aY_t^{*2} + (1-a)(C_{t+1}^* - C_t^*)^2 \right] \rho^{t-1}$$

$$(12') \quad \text{subject to } Y_t^* = \phi_2 m_t + \phi_1 m_t^* - \phi_3 s_t$$

$$(13') \quad C_{t+1}^* - C_t^* = \eta_2 m_t + \eta_1 m_t^* - \eta_3 s_t$$

where

$$(14a) \quad s_{t+1} = cs_t + bm_t - bm_t^*$$

and

$$(14b) \quad m_t = f_t(s_t), \quad m_t^* = f_t^*(s_t^*),$$

and the minimizations in (10) and (10') are performed over the policy rules  $f_t$  and  $f_t^*$ , respectively, under different modes of decision making.

## II. Derivation of Noncooperative Equilibria

Equations (11)–(14) specify a dynamic game; the solution to which will be considered under different behavioral assumptions for policymakers in each country. Specifically, we will study the equilibrium solution under the assumption of (i) Cournot-Nash, (ii) Stackelberg behavior on the part of the policymakers.

To begin, we first substitute (12) and (13) into (11) and (12'), (13') into (11'), enabling us to express each country's objective function in terms of only the state variable,  $s_t$ , and the control variables of both countries,  $m_t, m_t^*$ . The resulting expressions are

$$(15) \quad J_T = \sum_{t=1}^T \left[ Q_1 s_t^2 + 2Q_2 s_t m_t + 2Q_3 s_t m_t^* + 2Q_4 m_t m_t^* + Q_5 m_t^2 + Q_6 m_t^{*2} \right] \rho^{t-1}$$

$$(15') \quad J_T^* = \sum_{t=1}^T \left[ Q_1^* s_t^{*2} + 2Q_2^* s_t m_t^* + 2Q_3^* s_t m_t + 2Q_4^* m_t m_t^* + Q_5^* m_t^{*2} + Q_6^* m_t^2 \right] \rho^{t-1},$$

where

$$Q_1 \equiv a\phi_3^2 + (1-a)\eta_3^2 \equiv Q_1^*;$$

$$Q_2 \equiv a\phi_1\phi_3 + (1-a)\eta_1\eta_3 \equiv -Q_2^*;$$

$$Q_3 \equiv a\phi_2\phi_3 + (1-a)\eta_2\eta_3 \equiv -Q_3^*;$$

$$Q_4 \equiv a\phi_1\phi_2 + (1-a)\eta_1\eta_2 \equiv Q_4^*;$$

$$Q_5 \equiv a\phi_1^2 + (1-a)\eta_1^2 \equiv Q_5^*;$$

$$Q_6 \equiv a\phi_2^2 + (1-a)\eta_2^2 \equiv Q_6^*.$$

Together with the evolution equation for the state variable, (14a), and the policy rules, (14b), the expressions for the cost functionals  $J_T$  and  $J_T^*$ , as given by (15), (15'), provide a convenient framework for the application of the available theory on dynamic games to this two-country model.<sup>8</sup>

#### A. Closed-Loop (Feedback) Nash Equilibrium Solution

The first type of equilibrium we will be addressing is the noncooperative Nash equilibrium under the so-called feedback information pattern (for both countries) as dictated by (13b).<sup>9</sup> Using the recursive technique given in Basar and Geert Olsder (1983, ch. 6), the solution of the dynamic game can be shown to be unique, and linear in the current value of the state, yielding the expressions given below in Proposition II.1. It is also time consistent.<sup>10</sup>

**PROPOSITION II.1:** *For the T-period dynamic game, the feedback Nash equilibrium solution is unique and is given by*

$$(16) \quad m_t = f_{t,T}(s_t) = \alpha_\tau s_t$$

$$(16') \quad M_t^* = f_{t,T}^*(s_t) = \alpha_\tau^* s_t$$

$$\tau \equiv T - t; \quad t = 1, 2, \dots, T,$$

where

$$(17) \quad \alpha_\tau = \frac{q_{2,\tau} q_{5,\tau}^* - q_{2,\tau}^* q_{4,\tau}}{q_{4,\tau} q_{4,\tau}^* - q_{5,\tau} q_{5,\tau}^*}$$

$$(17') \quad \alpha_\tau^* = \frac{q_{2,\tau}^* q_{5,\tau} - q_{2,\tau} q_{4,\tau}^*}{q_{4,\tau}^* q_{4,\tau} - q_{5,\tau}^* q_{5,\tau}}$$

<sup>8</sup>The details of the proofs are contained in a longer version of this paper; see T. Basar, Turnovsky, and V. d'Orey (1985).

<sup>9</sup>In the one-period game, this equilibrium reduces to the usual Cournot equilibrium.

<sup>10</sup>All solutions, being based on feedback rules, which are determined by using dynamic programming methods, are time consistent.

$$(18) \quad q_{1,\tau} = \rho c^2 \varepsilon_{\tau-1} + Q_1,$$

$$q_{1,\tau}^* = \rho c^2 \varepsilon_{\tau-1}^* + Q_1^*$$

$$q_{2,\tau} = \rho cb \varepsilon_{\tau-1} + Q_2,$$

$$q_{2,\tau}^* = -\rho cb \varepsilon_{\tau-1}^* + Q_2^*$$

$$q_{3,\tau} = -\rho cb \varepsilon_{\tau-1} + Q_3,$$

$$q_{3,\tau}^* = \rho cb \varepsilon_{\tau-1}^* + Q_3^*$$

$$q_{4,\tau} = -\rho b^2 \varepsilon_{\tau-1} + Q_4,$$

$$q_{4,\tau}^* = -\rho b^2 \varepsilon_{\tau-1}^* + Q_4^*$$

$$q_{5,\tau} = \rho b^2 \varepsilon_{\tau-1} + Q_5,$$

$$q_{5,\tau}^* = \rho b^2 \varepsilon_{\tau-1}^* + Q_5^*$$

$$q_{6,\tau} = \rho b^2 \varepsilon_{\tau-1} + Q_6,$$

$$q_{6,\tau}^* = \rho b^2 \varepsilon_{\tau-1}^* + Q_6^*$$

and

$$(19) \quad \varepsilon_\tau = q_{1,\tau} + 2q_{2,\tau} \alpha_\tau + 2q_{3,\tau} \alpha_\tau^* + 2q_{4,\tau} \alpha_\tau \alpha_\tau^* + q_{5,\tau} \alpha_\tau^2 + q_{6,\tau} \alpha_\tau^{*2}$$

$$(19') \quad \varepsilon_\tau^* = q_{1,\tau}^* + 2q_{2,\tau}^* \alpha_\tau^* + 2q_{3,\tau} \alpha_\tau + 2q_{4,\tau}^* \alpha_\tau \alpha_\tau^* + q_{5,\tau} \alpha_\tau^{*2} + q_{6,\tau}^* \alpha_\tau^2$$

$$\tau = 0, 1, 2, \dots, T-1$$

with the boundary conditions for the  $q$ 's and  $q^*$ 's being

$$q_{i,0} = Q_i, \quad q_{i,0}^* = Q_i^*$$

$$i = 1, 2, \dots, 5$$

The corresponding Nash equilibrium values for  $J_T$  and  $J_T^*$ , denoted by  $J_{T,1}$ ,  $J_{T,1}^*$ , respectively, are

$$(20) \quad J_{T,1} = \varepsilon_{T-1} s_1^2, \quad J_{T,1}^* = \varepsilon_{T-1}^* s_1^{*2}.$$

Note that the unique Nash equilibrium optimal policy rules, characterized by the two sequences  $\{\alpha_\tau\}$  and  $\{\alpha_\tau^*\}$ , depend only on  $\tau$ , the difference between the terminal time  $T$  and the current time  $t$ , and which therefore represents the "time to go." Since the problem is time invariant, this implies that letting  $T \rightarrow \infty$  is equivalent to letting  $\tau \rightarrow \infty$ , in the determination of the stationary equilibrium policy rules.

The time paths followed by the two economies are obtained by substituting (16) and (17) into (7)–(9). A question which our analysis leaves unresolved concerns  $s_1$ , the initial real exchange rate. In rational expectations models this is often determined through an initial jump, which takes the economy onto a stable manifold, thereby ensuring convergence. With active stabilization, however, convergence can be attained without such jumps, as long as the system is controllable. This condition is obviously met for the present model and indeed our numerical results below confirm stability for all parameter sets yielding nondegenerate optimization problems. Under these circumstances, the motivation for the initial jump is not apparent.<sup>11</sup>

No violation either to stability, or to the rationality of expectations, is incurred by treating  $s_1$  as being determined by past monetary policies. Furthermore, other than as a scale factor, the determination of  $s_1$  is irrelevant to either the nature of the optimal feedback policy rules, or the time profile of the dynamics, which are our main concern. On the other hand, it would be straightforward, and lead to little change if one allowed  $s_1$  to be determined by some tradeoff of the future costs contained in (20), with some initial adjustment costs.<sup>12</sup>

<sup>11</sup>In effect we have the familiar indeterminacy problem arising from having "too many" stable roots. Using a higher-dimension system, Currie and Levine (1985) handle the issue of jumps by invoking the notion of controllability and considering solutions which have the saddlepoint property. Our procedure is the one-dimension analogue of this.

<sup>12</sup>An example of this for optimal monetary policy in an  $s$  small open economy is given by P. J. Stemp and Turnovsky (1987).

## B. Feedback Stackelberg Solution

The Nash equilibrium solution considered above is a symmetric equilibrium concept in terms of the roles of the players in the game. Suppose now, that one of the two policy-makers (called the leader) has the power to dominate the decision process. Even in a model such as this, where the structures of the two economies are taken to be symmetric, this case is of interest. For example, Country 1 can be taken to be the United States, say, and Country 2 to be comprised of a large number of small countries, which together make up Europe and collectively are approximately equivalent to the United States in size and structure. Nevertheless, it does not seem unreasonable to assume that the United States by consisting of a single decision-making unit, is able to be the dominant player. Such an asymmetry in the roles of the players leads to the Stackelberg solution, which admits two different definitions—global and feedback—depending upon whether he can enforce his policy over the entire duration of the game or only from one period to another. The latter mode of play, which corresponds to the "feedback Stackelberg solution," allows for a recursive derivation and is the equilibrium solution we will adopt.

In the derivation of the feedback Stackelberg solution, we follow the recursive technique presented in Basar and Olsder (1983, ch. 7), which is parallel to the derivation of the Nash solution, the only difference being that now at every stage a static Stackelberg game is solved, instead of a Nash game. Taking Country 1 as the leader and Country 2 as the follower, the solution to the  $T$ -period dynamic game (which, like the feedback Nash solution, is time consistent) is presented as follows:

**PROPOSITION II.2:** *For the  $T$ -period dynamic game, the feedback Stackelberg solution is unique and is given by*

$$(21) \quad m_t = f_{t,T}(s_t) = \alpha_\tau s_t,$$

$$(21') \quad M_t = f_{t,T}^*(s_t) = \alpha_\tau^* s_t,$$

$$\tau \equiv T - t; \quad t = 1, 2, \dots, T,$$

where

$$(22) \quad \alpha_\tau = \frac{q_{2,\tau} - q_{3,\tau} \frac{q_{4,\tau}^*}{q_{5,\tau}^*} - q_{4,\tau} \frac{q_{2,\tau}^*}{q_{5,\tau}^*} + q_{6,\tau} \frac{q_{2,\tau}^*}{q_{5,\tau}^*} \frac{q_{4,\tau}^*}{q_{5,\tau}^*}}{\frac{2q_{4,\tau} q_{4,\tau}^*}{q_{5,\tau}^*} - q_{5,\tau} - q_{6,\tau} \left( \frac{q_{4,\tau}^*}{q_{5,\tau}^*} \right)^2}$$

$$(22') \quad \alpha_\tau^* = -\frac{1}{q_{5,\tau}^*} [q_{2,\tau}^* + q_{4,\tau}^* \alpha_\tau]$$

and  $q_{i,\tau}$ ,  $q_{i,\tau}^*$ ,  $i=1,2,\dots,5$ ,  $\varepsilon_\tau$ ,  $\varepsilon_\tau^*$  satisfy the same equations as before, that is, (18), (19), and (19'). The corresponding feedback Stackelberg equilibrium values for  $J_{T,1}$  and  $J_{i,1}^*$  are

$$J_{T,1} = \varepsilon_{T-1} s_1^2, \quad J_{T,1}^* = \varepsilon_{T-1}^* s_1^2.$$

### III. Cooperative Equilibrium

Suppose now that the two players agree to cooperate by minimizing their joint cost function,  $\tilde{J}_T = J_T + J_T^*$ . By substitution, this can be written as

$$(23) \quad \text{Min } \tilde{J}_T = \sum_{i=1}^T [\tilde{Q}_1 s_i^2 + 2\tilde{Q}_2 s_i m_i + 2\tilde{Q}_3 s_i m_i^* + 2\tilde{Q}_4 m_i m_i^* + \tilde{Q}_5 m_i^2 + \tilde{Q}_6 m_i^{*2}] \rho^{t-1}$$

subject to (13a), (13b), where

$$\tilde{Q}_1 = Q_1 + Q_1^*; \quad \tilde{Q}_4 = Q_4 + Q_4^*$$

$$\tilde{Q}_2 = Q_2 + Q_2^*; \quad \tilde{Q}_5 = Q_5 + Q_5^*$$

$$\tilde{Q}_3 = Q_3 + Q_3^*; \quad \tilde{Q}_6 = Q_6 + Q_6^*,$$

and  $Q_i$ ,  $Q_i^*$  are defined previously. This is a standard problem in intertemporal optimization, the solution to which is

$$(24a) \quad m_i = \tilde{\alpha}_\tau s_i$$

$$(24b) \quad m_i^* = \tilde{\alpha}_\tau^* s_i$$

where

$$(25a) \quad \tilde{\alpha}_\tau = \frac{\tilde{q}_{3,\tau} \tilde{q}_{4,\tau} - \tilde{q}_{2,\tau} \tilde{q}_{6,\tau}}{\tilde{q}_{5,\tau} \tilde{q}_{6,\tau} - \tilde{q}_{4,\tau}^2}$$

$$(25b) \quad \tilde{\alpha}_\tau = \frac{\tilde{q}_{2,\tau} \tilde{q}_{4,\tau} - \tilde{q}_{3,\tau} \tilde{q}_{5,\tau}}{\tilde{q}_{5,\tau} \tilde{q}_{6,\tau} - \tilde{q}_{4,\tau}^2},$$

and  $\tilde{q}_{i,\tau}$  are determined by equations analogous to (18) and (19). The corresponding cooperative equilibrium value for  $\tilde{J}_{T,1}$  is

$$\tilde{J}_{T,1} = \tilde{\varepsilon}_{T-1} s_1^2,$$

with the costs being borne equally by two economies.

### IV. Stationary Equilibria

The solutions discussed in Sections III and IV are based on a finite time horizon. These games are expressed in recursive intensive form, which enables the steady-state equilibrium solutions to be derived as the limit of the iterative solutions given above. This procedure is discussed in an expanded version of the paper. Here we only refer briefly to some of the issues.

The stationary solutions are obtained by considering the limits, as  $\tau \rightarrow \infty$ , of the solutions given in (17), and (18), in the case of the Nash game, and the analogous equations for the other games. Defining the limits

$$\lim_{\tau \rightarrow \infty} \alpha_\tau = \bar{\alpha}, \quad \lim_{\tau \rightarrow \infty} \alpha_\tau^* = \bar{\alpha}^*,$$

and likewise for  $\bar{q}$ ,  $\bar{q}^*$ ,  $\bar{\varepsilon}$ ,  $\bar{\varepsilon}^*$ , The steady-state policy rules are given by

$$(26a) \quad m_i = \bar{\alpha} s_i$$

$$(26b) \quad m_i^* = \bar{\alpha}^* s_i,$$

where  $\bar{\alpha}$ ,  $\bar{\alpha}^*$ , are the solutions to the set of equations obtained by letting  $\tau \rightarrow \infty$  in (18), and (19). These constitute highly nonlinear coupled algebraic equations in  $\bar{q}$ ,  $\bar{\varepsilon}$ , and the recursive procedure we have outlined provides a solution (which is unique) to these coupled equations.

The equilibrium steady-state path, obtained by substituting (26a), (26b) into (9), is given by

$$s_{t+1} = (c + b(\bar{\alpha} - \bar{\alpha}^*))s_t \equiv \theta s_t, \\ t = 1, 2, \dots,$$

from which stability follows if and only if

$$(27) \quad |c + b(\bar{\alpha} - \bar{\alpha}^*)| < 1.$$

The parameter  $\theta$  is the steady-state rate of convergence and governs the rate of convergence of all variables in the two economies. Our numerical simulations indicate that this condition is satisfied by our equilibrium solution candidates.

$$\lim_{T \rightarrow \infty} J_{T,1} = \bar{e}s_1^2 \quad \lim_{T \rightarrow \infty} J_{T,1}^* = \bar{e}^*s_1^{*2}$$

both of which are finite.

## V. Numerical Procedures

The parameters describing the optimal policies under the various strategic regimes are themselves complex functions of the underlying parameters of the model. Thus apart from revealing the general nature of the optimal policies, it is difficult to gain much insight into the general welfare implications of the different regimes. We therefore use plausible numerical parameter values to evaluate the rules and their welfare differences.

Table 1 indicates a set of base parameter values. These are chosen on the basis of reasonable empirical evidence. The elasticity of the demand for domestic output with respect to the foreign output is  $d_1 = .3$ , the semielasticity of the demand for output with respect to the real interest rate is  $d_2 = .5$ , while the elasticity of the demand for output with respect to the relative price is  $d_3 = 1$ . The income elasticity of the demand for money is  $e_1 = 1$ , while the semielasticity of money demand with respect to the nominal interest rate is .5. The share of domestic consumption is .6 for the two economies; the slopes of their respective Phillips curves are .75. The relative weights given to output

TABLE 1—PARAMETER VALUES<sup>a</sup>

### A. Base Set (Parameter Set 1)

$$d_1 = .3, \quad d_2 = .5, \quad d_3 = 1, \quad e_1 = 1.0, \quad e_2 = .5 \\ \delta = .6, \quad \gamma = .75, \quad \alpha = .75, \quad \rho = .9$$

### B. Variants (Parameter Sets 2–28)

$$d_1: 0, .2, .4, .6, .8 \\ d_2: .01, .25, 1.0, 10 \\ e_1 = 0, 0.5 \\ e_2 = .1, 1.0, 10 \\ \delta = .5, .75, .99 \\ \gamma = .5, 1 \\ \alpha = 1, .2, .4, .6, 1 \\ \rho = .8, 1, \quad \gamma \text{ (static)}$$

<sup>a</sup>Since  $d_3$  appears as a scale variable applied to  $s$ , the results are insensitive to changes in  $d_3$  (except for a scale factor). We therefore do not consider changes in  $d_3$ , but instead have maintained  $d_3 = 1$  throughout.

stabilization in the objective function is  $\alpha = .75$ , while the discount rate is  $\rho = .9$ .

While these values seem reasonable, they are uncertain. In Part B of the table we therefore consider variants of these values, allowing the parameters to range below low values and high values. Note that since  $d_2$ ,  $e_2$ , are semielasticities the values of  $d_2 = .5$ ,  $e_2 = .5$ , correspond to elasticities of around .03, .05, respectively.<sup>13</sup>

To consider all combinations of these parameter values would be impractical. Our approach, therefore, is to begin with the base set and introduce one parameter change at a time. Performing these changes in  $d_1$ ,  $d_2$ ,  $e_1$ ,  $e_2$ ,  $\delta$ ,  $\gamma$ ,  $\alpha$ , and  $\rho$  gives a total of 28 parameter sets. Parameter set 1 is the base set; sets 2–28 are obtained by substituting the corresponding values into the base set.

## VI. Alternative Equilibria: Base Parameter Set

Figures 1–4 illustrate the time paths for the equilibrium solutions corresponding to the base parameter set, described in Table 1. These have been drawn for an initial unit

<sup>13</sup>These statements are based on values of  $I = .10$ ,  $\dot{P} = .04$ . Larger values are considered in the sensitivity analysis.



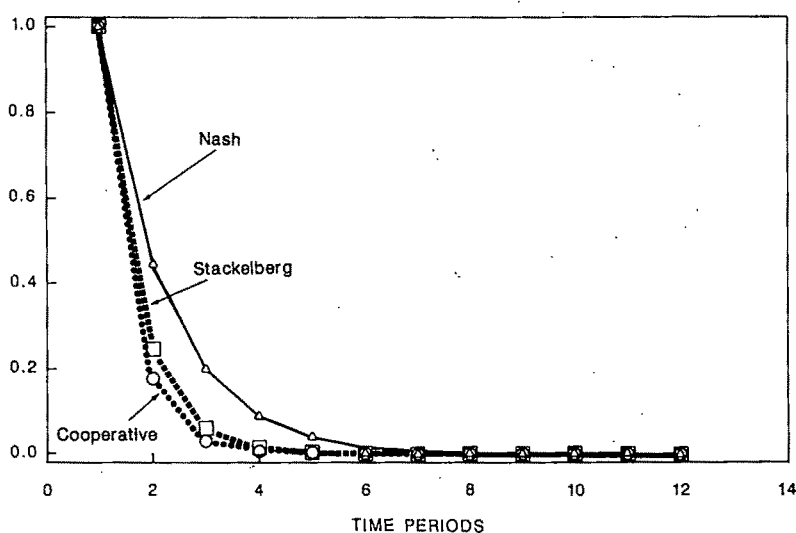


FIGURE 1. TIME PATHS FOR REAL EXCHANGE RATE

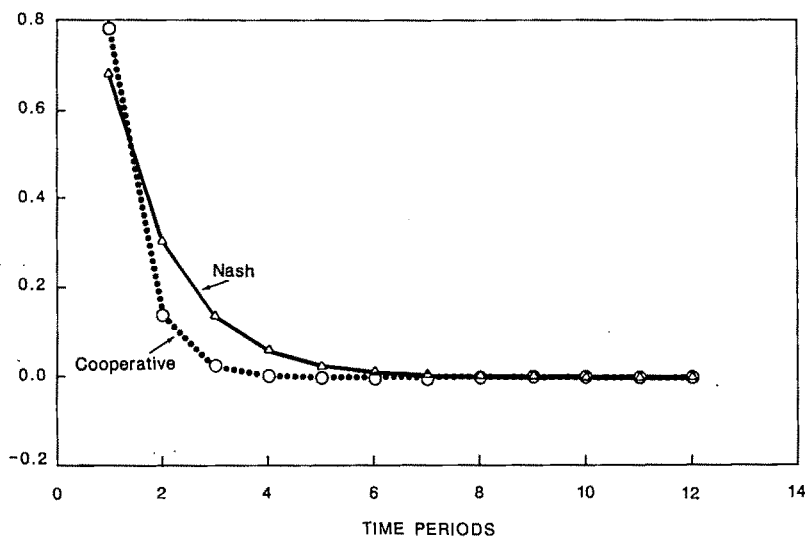


FIGURE 2A. TIME PATHS FOR REAL MONEY SUPPLY

positive shock in the relative price  $s$ , that is, for a given initial real depreciation of the currency of Country 1. The figures are drawn for a time horizon of  $T=12$  periods. The three equilibrium solutions are discussed in turn.

#### A. Feedback Nash

The time paths for  $s_t$ ,  $m_t$ ,  $Y_t$ , and  $\Delta C_{t+1} \equiv (C_{t+1} - C_t)$  under feedback Nash behavior are illustrated in Figures 1, 2A-4A. Given The symmetry of the model, the effects on

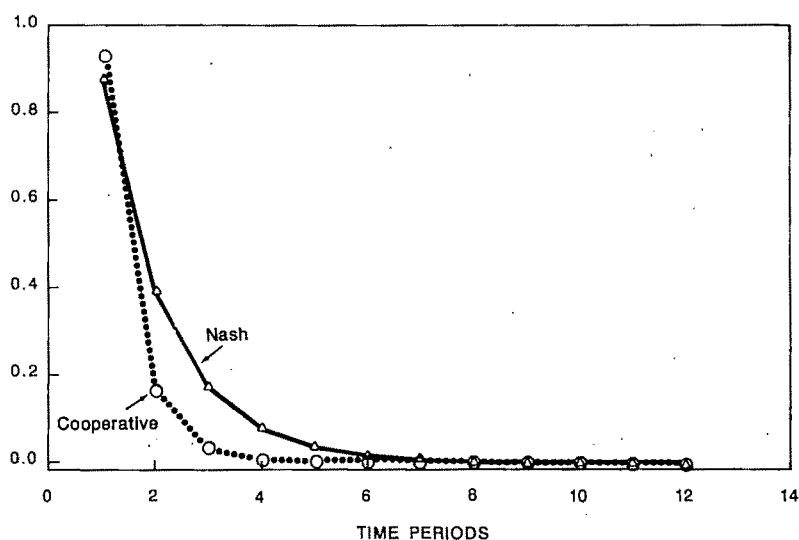


FIGURE 3A. TIME PATHS FOR REAL OUTPUT

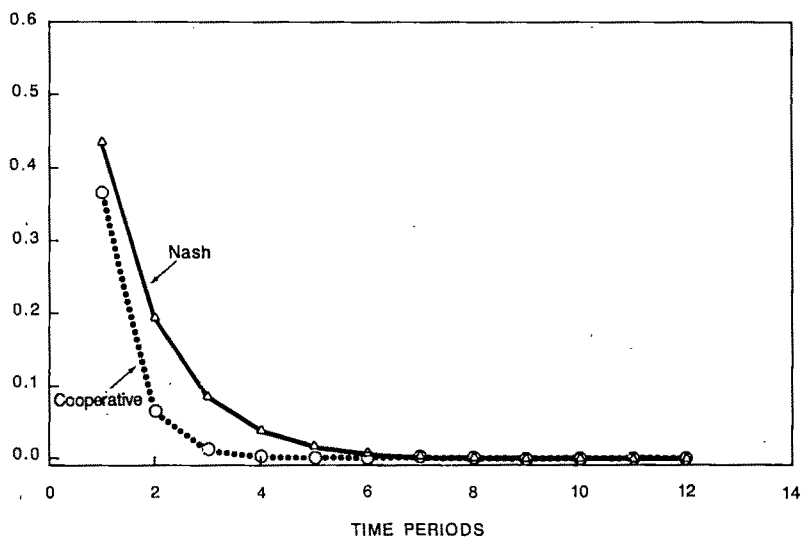


FIGURE 4A. TIME PATHS FOR CONSUMER PRICE INDEX

the two economies are identical (in magnitude), so that the time paths for  $m^*$ ,  $Y^*$ ,  $\Delta C^*$  are just mirror images of those of  $m$ ,  $Y$ , and  $\Delta C$ .

As a benchmark, suppose initially that in response to a unit increase in  $s$ , there is no response on the part of the two policymakers; that is,  $m_t = m_t^* = 0$ . In effect, the

policymakers agree to allow the exchange rate to float freely, so that this is a kind of cooperative equilibrium. In the first instance, the positive disturbance in  $s$  raises the demand for domestic output and reduces the demand for foreign output. This leads to an increase in domestic output  $Y$ , matched by an equivalent decrease in foreign output  $Y^*$ .

With the real money stocks held constant in both economies, these changes in output will lead to an increase in the domestic interest rate, accompanied by a decrease in the foreign interest rate, the net effect of which is to cause the rate of exchange depreciation of the domestic currency to increase. The increase in domestic output leads to a rise in the inflation rate of domestic output. This together with the increase in the rate of exchange depreciation, causes the rate of inflation of the overall domestic CPI to increase; the opposite occurs abroad.

Next, suppose that as in Turnovsky-d'Orey (1986), each policymaker follows a Nash strategy, using a one-period (static) objective function. In this case, if Country 1 responds to the increase in the relative price by reducing its real money stock, this mitigates the expansion in domestic output, while at the same time raising the domestic interest rate. The opposite effects occur abroad, causing the rate of exchange depreciation of the domestic economy to increase relative to the benchmark case of a perfectly flexible regime. This in turn leads to larger short-run variations in the rate of inflation. The increases in welfare costs associated with this increased price variation more than offset the reduction due to lower income variation, causing the overall costs to increase. Note that this occurs despite the fact that the relative costs attached to output variation are greater. It is a reflection of the quadratic nature of the cost function which penalizes large variations more than proportionately.

Thus for Parameter set 1, Turnovsky-d'Orey (1986) demonstrate that the simple rule of essentially no intervention can dominate other forms of strategic behavior, including Nash and other equilibria. These findings are, however, parameter sensitive, as they note. Moreover, neither the absence of intervention, nor the optimal short-run Nash policy of leaning against the wind is desirable from the viewpoint of long-run welfare maximization. Both strategies are associated with large increases in the rate of exchange depreciation of the domestic currency (larger in the latter case), contributing to large increases in the real exchange rate, which in turn cause the fluctuations in outputs and

inflation in the two countries to increase over time. The repetition of either strategy in each period causes the real exchange rate  $s$  to follow a divergent time path, with welfare costs ultimately increasing without limit.

By contrast, the optimal Nash policy, which minimizes the intertemporal cost function, calls for precisely the *opposite* response, namely an initial real monetary expansion in Country 1, accompanied by a corresponding contraction in Country 2. These policies cause the level of output in Country 1 to now increase by more than it did in the benchmark situation. By the same token, and by the above reasoning, this causes the rate of exchange depreciation of the domestic currency, and hence the overall rate of domestic inflation, to decrease relative to the benchmark policy. Precisely the opposite effects occur abroad. The reduction in the domestic nominal rate of exchange depreciation, combined with the above movements in domestic and foreign outputs, leads to a reduction in the real exchange rate,  $s$ . This in turn leads to a mitigation in the fluctuations in outputs and inflation. As a result of implementing the second, and subsequent, stages of the optimization, the real exchange rate follows a convergent path, with steadily declining welfare costs.

For the twelve-period horizon illustrated in Figure 2A, the coefficients of the optimal policy rules  $\alpha$ ,  $\alpha^*$ , evolve as follows:

$$\alpha_\tau = -\alpha_\tau^* = .6847 \quad \tau = 11, \dots, 4$$

$$\alpha_3 = -\alpha_3^* = .6856$$

$$\alpha_2 = -\alpha_2^* = .6873$$

$$\alpha_1 = -\alpha_1^* = .6439$$

$$\alpha_0 = -\alpha_0^* = -.9036$$

The interesting point to observe is that in the last period, the policy rule switches sign. This reflects the change in optimal behavior in going from a static to an intertemporal objective function. In fact, the static analysis of Turnovsky-d'Orey (1986) is identical to the one-period-to-go solution of the present dynamic analysis. It is also of interest to

note that the policy rule converges to its steady state ( $\bar{\alpha} = .6837, \bar{\alpha}^* = .6837$ ), within just five periods. Finally, the speed of the adjustment of the economy along the optimal trajectory is given by  $\theta = .446$ , implying that around 55 percent of the adjustment is completed within the first period.<sup>14</sup>

The contrast between the optimal short-run and the optimal long-run policies is striking. We shall restrict our comments to the domestic economy, although analogous reasoning applies abroad. The basic cause of the difference stems from the *intratemporal* tradeoff between output and price variations incorporated in the model, and how this is shifted over time by the chosen policies. Under our assumptions, a depreciation of the domestic real exchange rate generates an increase in the demand for domestic output, leading to an increase in domestic output itself. In effect there is an outward shift in the domestic IS curve, which also leads to an increase in the domestic interest rate, while the increase in domestic output leads to increases in the inflation rates of both the price of domestic output and the domestic CPI. A domestic monetary contraction, as dictated by the short-run optimal strategy, mitigates the short-run fluctuations in output. But, at the same time, this causes the domestic interest rate to increase further, leading to additional increases in the rates of depreciation of both the nominal and real domestic exchange rates, the following period. This in turn leads to a further outward shift in the domestic IS curve and to a deterioration in the next period's tradeoff between output and inflation. For a myopic government, concerned only with the present, this longer-run adverse movement is irrelevant. The short-run contractionary policy, with its dampening effect on output, is clearly desirable.

However, the longer-run effects of the depreciation of the real exchange rate stemming from such a contraction are clearly destabilizing. As  $s_t$  continues to increase, longer-run fluctuations in real output are generated and variations in the inflation rate are increased. This is not in the interests of a government having a longer-run horizon. Instead, such a government will find it optimal to expand its money supply in the short run. While this will increase the short-run fluctuations in output, it will also stabilize the fluctuations in the real exchange rate, both in the short run and over time. As a consequence of this, a stable long-run adjustment path will be followed. It is interesting to note that this switch in policy occurs within just two periods, the minimum within which the intertemporal (in addition to the intratemporal) tradeoff is introduced. Moreover, the result is robust across all parameter sets.

### B. Feedback Stackelberg

At each stage, the follower's response to the leader's action is given by the relationship

$$m_t^* = -[q_{2,\tau}^* s_t + q_{4,\tau}^* m_t] / q_{5,\tau}^*.$$

This defines the follower's reaction function, the slope of which is  $-q_{4,\tau}^* / q_{5,\tau}^*$ . Being a function of  $\tau$ , this changes at each stage. Using the base parameter set, Turnovsky-d'Orey (1986) show that for the one-period objective,  $q_4^* = .046$ ,  $q_5^* = .305$ , so that the short-run reaction curve has a negative slope equal to  $-.15$ . This means that the foreign (follower) economy responds to a unit expansion in the domestic (leader) real money supply, with a monetary contraction of .15. Turnovsky-d'Orey characterize the negative slope as being a beggar-thy-neighbor world.<sup>15</sup>

<sup>14</sup>The feedback CCV results are generally similar to the feedback Nash. The main difference is that since both policymakers are aware of the other's actions, each moderates his own adjustment relative to Nash. For parameter set 1,  $\theta = .789$ , implying a substantially slower rate of adjustment.

<sup>15</sup>Note that this term is being used in a somewhat different way from its conventional usage. More commonly, a beggar-thy-neighbor world is one in which an expansionary policy in one country causes a contraction (in activity) abroad. Here we are using the term to characterize the interdependence between the adjust-

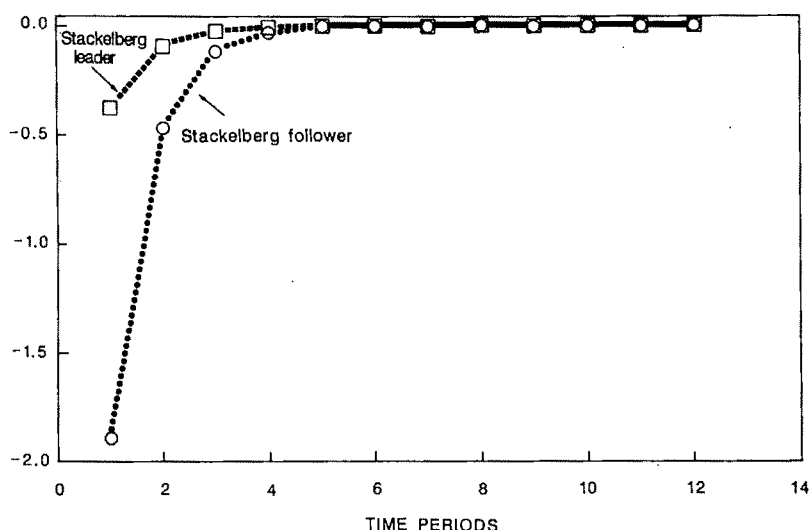


FIGURE 2B. TIME PATHS FOR REAL MONEY SUPPLY

This less than proportionate response by the follower implies that the Stackelberg equilibrium lies at a point on the follower's reaction function, away from the Nash equilibrium, in the direction of the follower's Bliss point. At this equilibrium point, the leader experiences a somewhat larger increase in output, accompanied by a smaller increase in inflation, relative to the Nash equilibrium, while for the follower, the negative fluctuations in both these variables are diminished in magnitude. Furthermore, while the welfare of the leader is higher than at the Nash equilibrium, the gains to the follower are relatively larger.

In the short-run Stackelberg equilibrium of the Turnovsky-d'Orey (1986) analysis, the real depreciation of the domestic currency leads to a monetary contraction by the leader. This action, together with the initial real appreciation of the foreign currency, has adverse effects on the level of output in the foreign economy. The foreign monetary authority (the follower) reacts to these nega-

tive effects by expanding its money supply, thereby tending to stabilize its level of output. For the same reasons as those given for the feedback Nash solution presented above, these responses lead to an increase in the relative price  $s_{12}$ , and cause the economy to embark on an unstable time path.

As shown in Figure 2B above, the appropriate initial responses become very different with an intertemporal objective function. Both the leader and the follower should now contract their respective real money stocks, with the contraction by the follower being significantly greater than in the Nash feedback case. The reason for the difference stems from the changed nature of the follower's short-run reaction function. In the initial period, we find  $q_4^* = -4.481$ ,  $q_5^* = 4.883$ , so that the slope of the reaction function is now .927 and is *positive*; this is characterized as being a locomotive world.

The leader knows that if he follows the feedback Nash strategy of expanding the money supply, the follower will tend to respond in a similar fashion. This tends to exacerbate the fluctuations in output in both economies, although the more balanced adjustment means that it is likely to be accompanied by smaller fluctuations in the rate of exchange depreciation (which responds to

ments in the policy instruments in the two economies, which in turn involves the slopes of the reaction functions. The same results apply to our usage of the term "locomotive" introduced below.

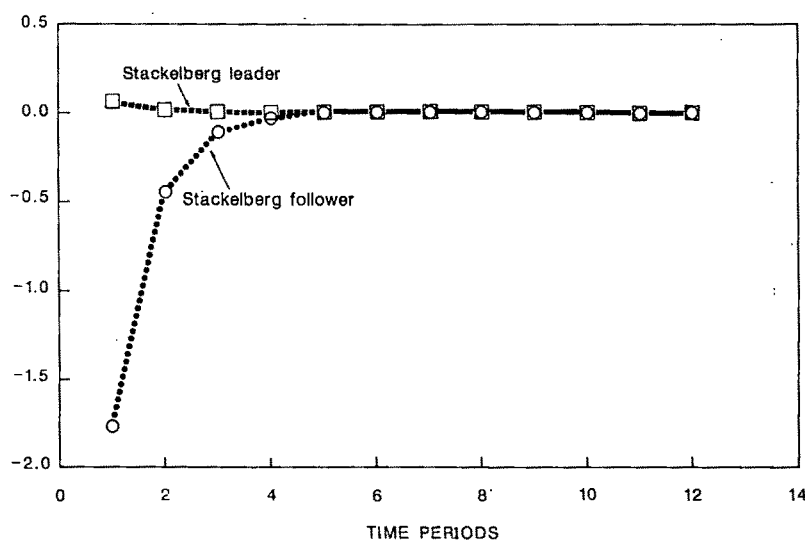


FIGURE 3B. TIME PATHS FOR REAL OUTPUT

*differential* monetary policies), and hence in the rate of inflation of the CPI. Given that the cost function assigns greater weight to output fluctuations than to fluctuations in inflation, this is a non-optimal situation, particularly for the leader. Accordingly, his strategy is to engage in a monetary contraction, thereby inducing an even greater contraction abroad by the follower. The fact that the contraction is relatively greater abroad causes an appreciation of the domestic currency, which in turn contributes to an appreciation of the real exchange rate, so that  $s_t$  begins to fall. This pattern of responses continues at each state, thereby enabling the economy to follow a stable path toward equilibrium.

A consequence of the initial worldwide monetary contraction is that the initial stimulating effects of the positive shift in the relative price  $s$  on the leader economy is largely eliminated. Indeed, in Period 1, output increases by only .06 units as compared to around .7 for the Nash equilibrium. At the same time, the monetary contraction means that the inflation of .4 percent under Nash becomes a deflation of .25 percent under Stackelberg. In the follower economy, the initial reductions in output and inflation under Nash are even greater under Stackel-

berg. These comparisons become evident upon examination of Figures 3A, 3B, and 4A, 4B.

Perhaps the most interesting feature of these results is the contrast in the welfares of the leader and follower between the single period and the multiperiod time horizon. We have already noted that for a one-period horizon, the follower is better off than the leader, with both being better off than under Nash. Now we see that over time, the leader improves his welfare vastly, though at the expense of the follower. The welfare costs under Nash to both are .759. Under feedback Stackelberg, however, the leader's costs are reduced to .020, making him much better off, while for the follower they rise to 2.758, resulting in a considerable loss in welfare.

The key to the difference between the short-run and long-run welfare costs is the switch in the follower's reaction function, which occurs over time. As noted, the optimal one-period policies, involving a combination of monetary contraction in the domestic economy and a monetary expansion abroad, destabilize the real exchange rate, affecting both countries adversely over time. To a myopic policymaker this is of no concern. But to a farsighted policymaker the long-run instability becomes important. The

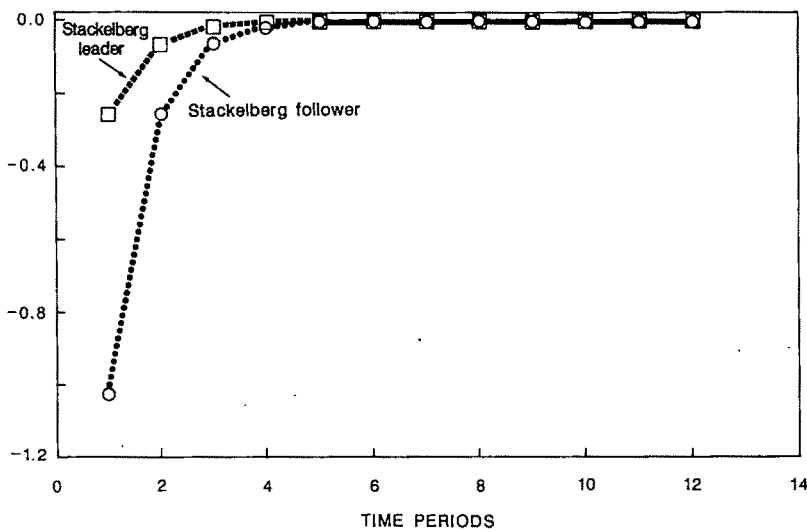


FIGURE 4B. TIME PATHS FOR CONSUMER PRICE INDEX

follower knows this and realizes that at each point of time it is up to him to respond in such a way to the leader's actions, to ensure that the relative price follows a stable path. To do this he responds to the real appreciation in his currency with a monetary contraction, at the same time trying to match more closely the qualitative response of the leader. The reason for this is that since  $s_t$  depends critically upon the difference in the real money stocks,  $(m_t - m_t^*)$ , minimizing this difference will tend to reduce the instability in  $s_t$ . Thus he will tend to contract when the leader contracts, and vice versa.

For his part, the leader knows the follower's response. But he also knows that the monetary contraction undertaken by the follower in response to his own actions will tend to have an adverse effect on his economy, and he therefore compensates by contracting less. In forcing the follower to respond at each stage to ensure that the relative price is stabilized, the leader is able to exploit his leadership more effectively over time. Basically he can act in his own self-interests and is able to impose most of the burden of adjustment on the follower, forcing him to bear the bulk of the adjustment costs. This reversal of the relative welfares occurs even within a two-period horizon, although

the differences increase with the length of the time horizon.

This finding raises serious questions of conflict in a multiperiod horizon. Obviously, in this situation neither country will agree to be the follower, raising serious doubts about the viability of the Stackelberg regime, unless there is some other mechanism whereby leadership is determined and enforced.

Finally, the convergence properties of the base parameter set can be summarized. For the 12-period horizon illustrated in Figures 2A and B the coefficients of the policy rules  $\alpha_\tau, \alpha_\tau^*$ , evolve as follows

$$\alpha_\tau = -.3717, \quad \alpha_\tau^* = -.1891, \quad \tau = 11, \dots, 5$$

$$\alpha_4 = -.3718, \quad \alpha_4^* = -.1890$$

$$\alpha_3 = -.3716, \quad \alpha_3^* = -.1881$$

$$\alpha_2 = -.3493, \quad \alpha_2^* = -.1778$$

$$\alpha_1 = -.1004, \quad \alpha_1^* = -.1130$$

$$\alpha_0 = -.7767, \quad \alpha_0^* = .8843$$

In this case, the convergence of the policy rule to its steady-state form takes 5 periods. Note again the big jump in the size of the

coefficients, between the second to last and last period. While the leader should always adopt a leaning against the wind policy, the response of the follower changes qualitatively during these two periods. The speed of the adjustment of the economy, as described by  $\theta$ , is .247, implying that 75 percent of the adjustment occurs within one period. This is considerably faster than for the feedback Nash equilibrium.

### C. Cooperative Equilibrium

Under noncooperative Nash behavior, the monetary authorities in both countries ignore the fact that their own policy responses to the initial disturbance in the real exchange rate have stabilizing effects abroad. For example, the monetary expansion in Country 1 causes both output and CPI inflation abroad to rise, thereby reducing the falls in these quantities abroad resulting from the combined effects of the initial disturbance together with the Nash response of the foreign monetary authority. The opposite applies with respect to Country 2. By taking these externalities into account, the cooperative equilibrium calls for a larger initial monetary expansion in Country 1, accompanied by an equivalently larger contraction in Country 2, relative to the Nash equilibrium. This exacerbates the short-run movements in output in the two economies, while reducing the relative movements in the interest rates. The rate of exchange depreciation of the domestic currency is reduced further (relative to Nash), thereby generating a smaller increase in the domestic rate of inflation. Again, precisely the opposite reactions occur abroad. The increased interventions by the two monetary authorities causes a substantial drop in the real exchange rate in the next period, which in turn causes reductions in domestic output and inflation. In fact, the rate of convergence of the cooperative equilibrium is so rapid, that even though output fluctuations are increased initially (relative to Nash), by the second period the relative price has been reduced to such a degree to cause the fluctuations in both output and inflation to be less than in the Nash equilibrium.

For the 12-period horizon, the coefficients of the optimal policy rules  $\alpha_\tau$ ,  $\alpha_\tau^*$ , follow

$$\alpha_\tau = -\alpha_\tau^* = .7859 \quad \tau = 11, \dots, 3$$

$$\alpha_2 = -\alpha_2^* = .7847$$

$$\alpha_1 = -\alpha_1^* = .7433$$

$$\alpha_0 = -\alpha_0^* = -.1410$$

As in the other cases, the policy rule switches sign in the last period. The convergence of the control law to its steady-state rule is even faster than before, occurring within just three periods. Also, as noted, the speed of adjustment of the system along the optimal trajectory is extremely fast, with around 82 percent of the adjustment occurring within the first period.

### D. Overview

These results show that, at least in the case of Parameter set 1, all three types of equilibria suggest a sharp contrast between optimal policy with a single-period objective and optimal policy within a dynamic objective function. Basically, the static analysis called for a monetary contraction for Country 1 (experiencing the positive shock in  $s$ ), accompanied by an equivalent monetary expansion in Country 2. These responses tend to reduce output fluctuations, while increasing fluctuations in inflation. Given the relative weights on these in the objective function, this is desirable for a one-period horizon. However, it is not optimal over the longer run. Such policies generate increasing fluctuations in the relative price, with increasing welfare costs in the future. These can be avoided by adopting policies which generate more variation in output and less variation in inflation.

Finally, we recall that Figures 1–4 have been drawn for 12 periods. This means that in the last period, the paths may begin to diverge, due to the myopic form of the policy rule in that period. Since, after 11 periods of optimal dynamic control,  $s$  is by then small,



such upturns may be imperceptible.<sup>16</sup> The time paths for the infinite horizon case are similar, except that the values of  $\alpha, \alpha^*$  are the same (at their steady-state values) for all periods.

## VII. Sensitivity Analysis

In order to determine the robustness of these results, we have recomputed the solutions across the 28 parameter sets discussed in Table 1. The last of these (set 28) is simply the one-period horizon considered by Turnovsky-d'Orey (1986), which corresponds to a discount rate of  $\rho = 0$ .<sup>17</sup> We consider the following three aspects summarizing the equilibria: (i) The steady-state policy rules,  $\bar{\alpha}, \bar{\alpha}^*$ ; (ii) the steady-state rate of convergence  $\theta$ ; (iii) the steady-state welfare costs.

### A. Steady-State Policy Rules

In virtually all cases, the Nash solution calls for leaning with the wind. Country 1 should expand its (real) money supply in response to the real depreciation of its currency; Country 2 should contract in response to the real appreciation of its currency. By contrast, the Stackelberg leader should almost always lean against the wind, while the follower should always do so with the exception of set 25 ( $\alpha = 1$ ), when the policy game degenerates.<sup>18</sup> Finally, except in polar cases, the cooperative equilibrium requires more intensive intervention than does the feedback Nash.

### B. Steady-State Rate of Convergence

For all but Parameter sets 10, 21, 25, and 28, all optimal paths converge. Parameter set

28 is the static case, which for reasons discussed at length always leads to divergence. Parameter sets 21 and 25 are the degenerate extremes, when the targets are always attained perfectly in each period. In this case, the divergence of  $s_t$  is irrelevant. It can always be accommodated by increasing adjustments in the controls  $m_t, m_t^*$ . The only genuine dynamic game in which divergence occurs is set 10, with  $d_2 = 10$ . This value violates the condition for a downward sloping IS curve and hence instability is not so surprising.

With the exception of the extreme set 7, for all other sets with  $\alpha = .75$ , the values of  $\theta$  under Nash, Stackelberg, and cooperative behavior,  $\theta_N, \theta_S$ , and  $\theta_C$ , respectively, satisfy

$$\theta_C < \theta_S < \theta_N,$$

implying a clear ranking in the rates of convergence; the cooperative equilibrium is faster than Stackelberg, which in turn is faster than Nash.<sup>19</sup>

### C. Steady-State Welfare Costs and the Gains from Cooperation

The pattern of welfare costs is also remarkably stable and gives rise to a clear ranking among the equilibria. With the exception of the degenerate cases (sets 21, 25) and the static case (set 28), the ranking of the different solutions obtained for the base parameter set extends to all other cases. The Stackelberg leadership is the best, while being a Stackelberg follower is the worst equilibrium. In between these extremes we find that the cooperative equilibrium dominates Nash. The welfare costs to the Stackelberg leader are remarkably stable across parameter sets and take him close to his Bliss point (zero costs). By contrast, the Stackelberg follower does extremely poorly, questioning the

<sup>16</sup>In the case of the CCV solution, however, the upturn is in fact quite marked.

<sup>17</sup>Tables presenting the detailed results of the sensitivity analysis are available from the authors.

<sup>18</sup>Note that set 21 ( $\alpha = 0$ ), set 25 ( $\alpha = 1$ ) give rise to degenerate policy games. This is because in either case, the objective function of each policymaker reduces to just one target and the two policy instruments  $m, m^*$  enable each to be attained perfectly. All solutions converge to the same with zero welfare costs being incurred.

<sup>19</sup>For almost all parameter sets the slowest rate of adjustment is achieved under feedback CCV equilibrium. The reason is that since each policymaker takes account of his rival's actions, this induces caution and gives rise to a more gradual adjustment.

viability of this regime, relative to the alternatives.<sup>20</sup>

Allowing for compensation, the Nash equilibrium is the preferred noncooperative equilibrium from an overall welfare viewpoint. However, for the base parameter set it still yields welfare losses which are approximately 8 percent greater than those for the cooperative equilibrium. The gains from cooperation are generally of this order of magnitude, and are mildly parameter sensitive. Overall, the robustness of these results for the dynamic game are in sharp contrast to the rankings obtained by Turnovsky-d'Orey (1986) for the static game, which for the same parameter sets were found to be extremely parameter sensitive.

### VIII. Conclusions

In this paper we have developed dynamic strategic monetary policies using a standard two-country macro model under flexible exchange rates. Two types of noncooperative equilibria have been considered, namely feedback Nash and feedback Stackelberg. In addition, these have been compared to the Pareto-optimal cooperative equilibrium.

The optimal policies have been obtained as feedback rules in which the real money supplies in the respective economies are adjusted to movements in the real exchange rate. Even for a simple model such as this, the derivation of the optimal policies is highly complex, particularly in the limiting case of an infinite time horizon. For this reason, much of our work has proceeded numerically. In carrying out our simulations, we have compared the results obtained from the present dynamic analysis with those obtained previously for the same simulation sets, but using a single-period time horizon.

Many of the specific conclusions of our analysis have been noted previously. At this point, several general conclusions are worth highlighting. First, the optimal policies were

found to yield convergence for all three equilibria, in the case of virtually all parameter sets. A clear ranking in the rate of convergence was obtained, cooperative behavior yields the fastest convergence, followed by the feedback Stackelberg, with feedback Nash being the slowest.

The results indicate a sharp contrast in both the optimal policies and welfare between the previous results obtained for the short-run time horizon and the present results for the long run, thereby suggesting the importance of intertemporal and intratemporal tradeoffs. As far as welfare is concerned, while in the short run the ranking of the equilibria is highly parameter sensitive, in the long run the rankings are remarkably robust across parameter sets. Specifically, in the long run we find among the noncooperative equilibria, the Stackelberg leader to be the preferred equilibrium, followed by feedback Nash, and Stackelberg follower. The superiority of the Stackelberg leader suggests that it takes time for him to be able to exploit his position. The welfare gains from cooperation over Nash are typically of the order of 6–10 percent. Although these are modest, they are certainly not negligible.<sup>21</sup>

While these results are suggestive and appearing promising, we should note at least two important limitations of our analysis. First, it is based on two symmetric economies, and while this is an obvious natural starting point, it clearly needs to be relaxed. Second, the model itself is simple in terms of minimizing the order of the dynamics; extensions in the direction of generating a richer model structure are also desirable, before the results obtained can be maintained with confidence.

<sup>21</sup> These numerical estimates of the gains from cooperation are similar in magnitude to those obtained by Taylor (1985).

<sup>20</sup> The CCV equilibrium always leads to higher welfare costs than Nash. This reflects the fact that CCV is associated with slower adjustment, thereby contributing to larger intertemporal welfare costs.

### REFERENCES

- Basar, T., "A Tutorial on Dynamic and Differential Games," *presented at the 7th An-*

- nual Conference on Economic Dynamics and Control*, London, June 1985.
- \_\_\_\_\_, and Olsder, G. J., *Dynamic Noncooperative Game Theory*, New York: Academic Press, 1983.
- \_\_\_\_\_, Turnovsky, S. J. and d'Orey, V., "Optimal Strategic Monetary Policies in Dynamic Interdependent Economics," presented at the 7th Annual Conference on Economic Dynamics and Control, London, June 1985.
- Bresnahan, T. F., "Duopoly Models with Consistent Conjectures," *American Economic Review*, December 1981, 71, 934-45.
- Canzoneri, M. and Gray, J. A., "Monetary Policy Games and the Consequences of Noncooperative Behavior," *International Economic Review*, October 1985, 26, 547-64.
- Currie, D. and Levine, P., "Macroeconomic Policy Design in an Interdependent World," in W. H. Buiter and R. C. Marston, eds., *International Economic Policy Coordination*, New York: Cambridge University Press, 1985.
- Dornbusch, R., "Expectations and Exchange Rate Dynamics," *Journal of Political Economy*, December 1976, 84, 1161-76.
- Fershtman, C. and Kamien, M. I., "Conjectural Equilibrium and Strategy Spaces in Differential Games," in *Optimal Control Theory and Economic Analysis*, Vol. 2, G. Feichtinger, ed., Amsterdam: North-Holland, 1985.
- Hamada, K., "A Strategic Analysis of Monetary Interdependence," *Journal of Political Economy*, August 1976, 84, 677-700.
- Hughes Hallett, A. J., "Non-Cooperative Strategies for Dynamic Policy Games and the Problem of Time Inconsistency," *Oxford Economic Papers*, November 1984, 56, 381-99.
- Jones, M., "International Liquidity: A Welfare Analysis," *Quarterly Journal of Economics*, February 1983, 98, 1-23.
- Kamien, M. I. and Schwartz, N. L., "Conjectural Variations," *Canadian Journal of Economics*, May 1983, 16, 191-211.
- Miller, M. H. and Salmon, M., "Policy Coordination and Dynamic Games," in W. H. Buiter and R. C. Marston, eds., *International Economic Policy Coordination*, New York: Cambridge University Press, 1985.
- Oudiz, G. and Sachs, J., "International Policy Coordination in Dynamic Macroeconomic Models," in W. H. Buiter and R. C. Marston, eds., *International Economic Policy Coordination*, New York: Cambridge University Press, 1985.
- Perry, M. K., "Oligopoly and Consistent Conjectural Variations," *Bell Journal of Economics*, Spring 1982, 13, 197-205.
- Phelps, E. S., "Phillips Curves, Expectations of Inflation, and Optimal Unemployment Over Time," *Economica*, August 1967, 24, 254-81.
- Rogoff, K., "Can International Monetary Policy Coordination Be Counterproductive?," *Journal of International Economics*, May 1985, 18, 199-217.
- Stemp, P. J. and Turnovsky, S. J., "Optimal Monetary Policy in an Open Economy," *European Economic Review*, July 1987, 31, 111-35.
- Taylor, J. B., "International Coordination in the Design of Macroeconomic Policy Rules," *European Economic Review*, June 1985, 28, 53-81.
- Turnovsky, S. J., "Monetary and Fiscal Policy Under Perfect Foresight: A Symmetric Two-Country Analysis," *Economica*, May 1986, 53, 139-157.
- \_\_\_\_\_, and d'Orey, V., "Monetary Policies in Interdependent Economies: A Strategic Approach," in Symposium on the Coordination of Economic Policies Between Japan and the United States, *Economic Studies Quarterly*, June 1986, 37, 114-33.

# Exchange Controls, Capital Controls, and International Financial Markets

By ALAN C. STOCKMAN AND ALEJANDRO HERNÁNDEZ D.\*

*This paper examines the effects of restrictions on international financial markets in a general-equilibrium, rational-expectations model of a two-country world. Taxes or quantitative controls on purchases of foreign currency and on the income from foreign assets reduce international trade in goods, lower ex post welfare in the country in which they are imposed, and affect nominal prices and exchange rate.*

This paper analyzes the effects on prices and resource allocation of taxes on international financial transactions. We employ a general-equilibrium, rational-expectations model of a two-country world economy to examine the connections between these taxes and portfolio allocations, trade, and prices. We examine the effects of taxes on purchases of foreign currency—which we call *exchange controls*—and differential taxes on the income from foreign interest-bearing assets—which we call *capital controls*.<sup>1</sup>

Financial markets in our model are not subject to any limitations other than these taxes and a restriction that guarantees a monetary equilibrium by preventing barter. The rapid development of domestic and international financial markets over the past several years makes the assumption of sophisticated financial markets more em-

pirically relevant than in the past, and raises questions like those examined here about the effects of taxes and other limitations on those markets. The effects of exchange controls, capital controls of various forms, and other similar restrictions have been studied by R. Cumby, 1984; M. Obstfeld, 1984; C. Adams and J. Greenwood, 1985, and Greenwood and K. Kimbrough, 1985, 1987. Adams and Greenwood demonstrated an equivalence between dual exchange rates and capital controls, while Greenwood and Kimbrough showed that there is a similarity between exchange controls and taxes on trade, and examined the effects of fiscal policies in the presence of capital and exchange controls. Previous work in this area, however, has generally made implicit assumptions that restrict the availability of asset markets beyond the limitations implied by the explicit taxes or controls. This paper shows that the effects of exchange and capital controls depend critically on the availability of international financial markets for assets with state-contingent real returns, in ways that have been largely overlooked. Uncertain *prospective* changes in taxes or controls on acquisitions of foreign currency affect portfolio allocations in such a way that, if these prospective changes subsequently occur, their effects on prices and resource allocation are different than if financial markets for state-contingent assets had been missing. These differences result from the attempts of households to self-insure against the risks of future changes in government policies. These

\*University of Rochester and NBER, University of Rochester, NY 14726. Thanks go to Michael Dotsey, Marvin Goodfriend, Bennett McCallum, Torsten Persson, Lars E. O. Svensson, and anonymous referee, and participants in workshops at the Institute for International Economic Studies, Stockholm, the Graduate Institute of International Studies, Geneva, the Federal Reserve Banks of Richmond and Cleveland, and the National Bureau of Economic Research. Stockman gratefully acknowledges support from the National Science Foundation.

<sup>1</sup>Alternatively, we could consider quantitative restrictions. It is well known that, ignoring dissipation of rents due to rent-seeking activities, any quantitative restrictions can be duplicated by a set of taxes. Our results could also apply to other restrictions such as dual exchange rates.

attempts at self-insurance occur through trade in assets whose real return is affected by changes in government policies. Households use these assets to allocate wealth optimally across alternative states of the world. If enough financial markets are available then wealth-redistribution effects of policies will be nil, even though the available assets are subject to taxation. The main effects of policy changes, then, are due to aggregate wealth effects and substitution effects. A. Stockman and H. Dellas, 1986, and Stockman, 1987, study the effects of tariffs and various fiscal policies in the presence of sophisticated international financial markets. The idea that attempts to self-insure against risks of changes in future policies alters both current behavior and the effects of those policies appears in Lucas, 1976, and related issues are discussed in C. Sims, 1982, 1985, and T. Cooley, S. LeRoy, and N. Raymon, 1984a, b.

The analysis in this paper is restricted to a discussion of the positive effects of changes in exchange and capital controls; it abstracts from a discussion of the determinants of these policies, for which a more sophisticated model would be necessary. We find that an increase in domestic taxes on acquisitions of foreign currency is likely to raise the domestic terms of trade and that, unless accompanied by a sufficiently large increase in nominal domestic prices, this increase in the terms of trade is accomplished through an appreciation of the domestic currency. The increase in the terms of trade reduces the domestic consumption of importables. These taxes "work" in the sense of altering the terms of trade, the value of domestic currency, and the level of imports. On the other hand, they fail to raise domestic consumption of exportables unless exportables and importables are direct substitutes in utility, in the sense that utility functions have a negative cross derivative. Even in this case, *ex post* utility is likely to fall in the domestic country and increase in the foreign country. The contemporaneous effect of an increase in the current tax rate on the terms of trade and consumption levels is independent of whether that increase signals a change in the conditional probability

distribution of future taxes. On the other hand, the contemporaneous effects of an increase in the current tax rate on nominal prices and the exchange rate depend on whether high current exchange and capital controls raise the conditional probability of controls in the future.

### I. Optimization Problems of Representative Households

We will examine a model with two countries, each with a representative risk-averse household that consumes two perishable goods,  $X$  and  $Y$ . These goods are endowed (supplied perfectly inelastically) to the countries. There is complete specialization in endowments, which follow a stochastic process with known distribution for all agents. Trade occurs because of the difference in endowments and because tastes are permitted to differ across countries. By convention, the domestic country exports  $X$ . Let  $(\bar{x}_t, \bar{y}_t)$  denote the endowment of  $X$  at time  $t$  to the domestic country and the endowment of  $Y$  at time  $t$  to the foreign country. We assume that all households are price takers who maximize discounted expected utility over an infinite horizon.

There are two moneys,  $M$  and  $N$  (the domestic and foreign currencies), which are introduced with cash-in-advance constraints; we also assume that sellers' currencies are used for all transactions. These constraints require purchases of goods each period to be financed with money held by households prior to receipt of income from current sales of goods (or dividends paid by firms from current receipts).<sup>2</sup>

There are complete, or at least Pareto-efficient, international asset markets except for the restriction that assets may not pay interest, principal, or dividends directly as goods: they may only pay moneys or other assets. If assets were permitted to pay interest

<sup>2</sup>See Stockman, 1980; R. Lucas, 1976, 1982; E. Helpman, 1981; Helpman and A. Razin, 1982, 1984; L. Svensson, 1985a, b; or Stockman and Svensson, 1987, for further discussion. The use of buyers' currencies is examined in Helpman and Razin, 1984.

as physical goods, then households could engage in complete contingent contracting that would eliminate any need for subsequent transactions; without transactions, there cannot be a transactions demand for money, and there would be no monetary equilibrium. This restriction is implied by the cash-in-advance constraints.

"Firms" are defined as the recipients of the endowments in each country. Shares of firms may be traded, and we normalize the number of shares in domestic firms at one per capita, using world population. The same normalization is made for shares in foreign firms.

During each period households visit asset markets where assets are traded, interest or dividend payments are made, and taxes are paid or transfers received. Households, who leave asset markets with portfolios that include money to finance subsequent expenditures, then visit product markets and purchase goods using money carried over from asset markets. The process then repeats itself, with firms paying as dividends, at subsequent asset markets, their money receipts from previous sales of goods. One can think of households as buying goods from vending machines—firms—that require money; the money then sits in machines—at the firms—until they are emptied when product markets close.<sup>3</sup>

Money supplies of each country are assumed to be fixed. The only government policy considered here is the proportional taxation of purchases of foreign currency and receipts of foreign currency from other sources, such as sales of goods abroad, dividends from foreign equities, and interest or principal from other foreign assets. Government policies are partly "anticipated" in the sense that households have rational expecta-

tions and know the model and the true probability distributions that govern policies. On the other hand, any specific policy is partly "unanticipated" in the sense that the probability of its adoption may be arbitrarily small. The actual pattern of taxes over time arises from the equilibrium of a political system that is not explicitly modeled here. The political equilibrium each period is subject to some uncertainty, in that households are not able to predict perfectly future policies. Households are assumed to have rational expectations regarding the formation of policy and the exogenous productivity shocks. The model permits policies to be correlated in any way over time and across countries. In order to avoid the additional notation (with little interesting economics) associated with corner solutions for some assets, the government is assumed to set the same tax rate  $\tau_t$  on all acquisitions of foreign currency, regardless of the source. These tax rates may change over time, and will be treated here as exogenous.<sup>4</sup> The foreign government sets tax rate  $\tau_t^*$ . Domestic (foreign) government revenue from taxes is assumed to be refunded through lump-sum transfers  $z_t$  (or  $z_t^*$ ) to domestic (foreign) residents.

The random vector

$$(1) \quad s_t = (\bar{x}_t, \bar{y}_t, \tau_t, \tau_t^*)$$

$$\in B \subset R^2 \times [0,1] \times [0,1]$$

describes the realization of endowments and tax rates at date  $t$ . Let  $S \equiv \{s: s = \{s_t\}_{-\infty}^{\infty}\}$  denote the set of double infinite sequences. The probability space in our model is  $(S, \zeta, \pi)$  where  $\zeta$  is the  $\sigma$ -algebra generated by cylinder subsets of  $S$  and  $\pi$  is a probability measure. Households attach probabilities

<sup>3</sup>The dating of time periods is completely arbitrary and does not affect the results. Also, as Svensson (1985a,b) has shown, asset trading can occur continuously in cash-in-advance models. Product markets are assumed to be open only at certain times: this plays the same role as explicit transactions costs in a Baumol-Tobin model and generates a positive demand for money.

<sup>4</sup>More generally, policy changes could be endogenous: if the model of policy formation is deterministic, then households would have perfect foresight on actual policy in a rational expectations model; if the model of policy determination involves some uncertainty (or limited information to households), then households would treat this uncertainty exactly as they treat uncertainty from technology shocks, etc., by treating actual policy as the outcome of a stochastic process.

not to a particular realization in a particular time period, but to whole sequences of events. This allows us to consider a very rich set of correlations of taxes and endowments at a point in time and over time.

Consider now a stochastic process  $\{\omega_t\}_{t=0}^\infty$  defined on  $(S, \mathcal{F}, \pi)$ . This process is a sequence of functions from  $S$  to a Euclidean space, and can represent either decision variables or prices. If two sequences  $s, s' \in S$ ,  $s \neq s'$ , have the same history up to a period  $t$ , that is,  $s_j = s'_j$  for  $j = \dots, t-1, t$ , then  $\omega_j(s) = \omega_j(s')$  for  $j = \dots, t-1, t$ . In other words, only the past and current realizations affect decision variables. The realized value at time  $t$  will be known by agents during period  $t$  asset markets, when taxes are paid.

Households have time-additive von Neumann Morgenstern preferences. The representative domestic household maximizes

$$(2) \quad \mu(x, y) = E_0 \sum_{t=0}^{\infty} \beta^t U(x_t, y_t) \\ = \sum_{t=0}^{\infty} \beta^t \int_s U(x_t(s), y_t(s)) \pi(ds).$$

Similarly, the representative foreign household maximizes

$$(3) \quad \mu^*(x^*, y^*) = E_0 \sum_{t=0}^{\infty} \beta^t U^*(x_t^*, y_t^*) \\ = \sum_{t=0}^{\infty} \beta^t \int_s U^*(x_t^*(s), y_t^*(s)) \pi(ds).$$

The utility functions  $U$  and  $U^*$  have the standard properties and exhibit risk aversion. Preferences may differ across countries, though we assume discount rates are identical. The only restriction we impose on  $U$  and  $U^*$  is that they satisfy

$$(4) \quad U_{12} U_{12}^* < \min(U_{11} U_{22}^*, U_{11}^* U_{22}),$$

where subscripts on  $U$  and  $U^*$  denote partial derivatives. This assumption would reduce to the concavity assumption if foreign and domestic households were identical in preferences and opportunities, and it is suffi-

cient though not necessary for some of the results in the next section.<sup>5</sup>

Domestic households have initial assets  $A_0$ . The initial assets of foreign and domestic households must sum to the values of equities and moneys, but any arbitrary international distribution of wealth is permitted. Let  $M$  and  $N$  denote the quantities of domestic and foreign moneys held at the close of asset markets, and  $P_M$  and  $P_N$  denote the prices of these moneys in terms of some numeraire.  $H$  and  $K$  denote quantities of equities in domestic and foreign firms held at the close of asset markets, acquired at prices  $P_H$  and  $P_K$  in terms of the numeraire.<sup>6</sup> Finally,  $B_t(s)$  and  $F_t(s)$  are purchases of contingent claims to domestic and foreign moneys delivered in state  $s$  at time  $t$ . We will refer to these assets as (state-contingent) bonds. They are purchased today at prices  $P_{B_t}(s)$  and  $P_{F_t}(s)$ . Dividends from foreign equities and deliveries of foreign moneys from these contingent bonds will be subject to future uncertain taxation.

The budget constraint faced by the domestic households at asset markets at date 0 is

$$(5) \quad A_0 = P_{M0} M_0 + P_{N0} (1 + \tau_0) N_0 \\ + P_{H0} H + P_{K0} K - P_{M0} z_0 \\ + \sum_{t=1}^{\infty} \int [P_{B_t}(s) B_t(s) \\ + P_{F_t}(s) F_t(s)] \pi(ds).$$

As mentioned above,  $z_0$  is the lump-sum

<sup>5</sup>Because there are infinitely many states of the world, care must be exercised in the way prices are defined. Let prices be a stochastic process  $\{P_t\}_0^\infty$ , a sequence of functions from  $S$  into  $R^2$ . If the probability measure  $\pi$  is nonatomic, the probability of a particular price would be zero. To avoid this, we define  $P_t(s)$  as a price at date  $t$ , given state  $s$ , that would be observed if  $s$  were to occur with probability one. Then  $P_t(s) \pi(ds)$  is the price that is actually observed.

<sup>6</sup>It is necessary in our model for households to alter their portfolios of equities over time, so we do not put time subscripts on  $H$  and  $K$ . Recall that equity supplies are each unity.

refund of the tax revenue,

$$(6) \quad z_0 = \tau_0 N_0 e_0,$$

where  $e_0$  is the nominal exchange rate of time 0

$$(7) \quad e_0 = P_{N0}/P_{M0}.$$

The exchange rate in state  $s$  at date  $t$  is

$$(8) \quad e_t(s) = P_{Nt}(s)/P_{Mt}(s),$$

where  $P_{Mt}(s)$  and  $P_{Nt}(s)$  are prices (in terms of the numeraire) of domestic and foreign money in state  $s$  at date  $t$ .

A few things should be noticed about equation (5). First, although we allow any initial distribution of wealth, we assume that the domestic money is initially held exclusively by domestic residents, and foreign money by foreigners. Therefore, all foreign currency acquired by the domestic resident is subject to the tax. Second, the shares of the domestic and foreign firms are not dated. There are two reasons for this. First, as we prove later, trade in equities is redundant if trade in other contingent assets is allowed. Second, as we will demonstrate, no future asset trades are required at all. Finally, note that purchases of foreign bonds are not taxed; only the income they generate (as interest or principal) is subject to taxation.

The domestic household is also constrained in its purchases of goods by

$$(9) \quad m_t(s) \equiv M_t(s) - p_t(s)x_t(s) \geq 0,$$

$$n_t(s) \equiv N_t(s) - q_t(s)y_t(s) \geq 0,$$

where  $p$  and  $q$  are nominal prices (in domestic and foreign currencies) of the goods  $X$  and  $Y$ . These require purchases of each good at date  $t$  to be financed by money on hand at the close of asset markets at date  $t$ .

A domestic firm is endowed with  $\bar{x}_t$  which it sells at date  $t$ ; it therefore earns  $p_t(s)\bar{x}_t$  to pay as dividends during asset markets at  $t+1$ . A fraction  $H$  of these dividends is received by the representative domestic household, which owns  $H$  equities in domestic firms. Households also receive money

payments from other assets, receive lump-sum refunds of tax revenues, and (possibly) carry over unspent money from previous product markets. So

$$(10) \quad M_t(s) = m_{t-1}(s) + p_{t-1}(s)\bar{x}_{t-1}H + B_t(s) + z_t(s)$$

$$N_t(s) = n_{t-1}(s)$$

$$+ [q_{t-1}(s)\bar{y}_{t-1}K + F_t(s)]/(1 + \tau_t),$$

where the transfer

$$(11) \quad z_t(s) = \tau_t [q_{t-1}(s)\bar{y}_{t-1}K + F_t(s)] \times P_{Ft}(s)/P_{Bt}(s)$$

is taken as given by the domestic household when it maximizes utility.<sup>7</sup>

The form of equations (10) implicitly assumes that taxes on foreign currency acquisitions are paid in units of foreign currency, while the lump-sum refund to domestic households is made in units of domestic currency. Therefore governments must participate in currency markets. This assumption is not important for any of our results.

The taxes on acquisitions of foreign currency may equivalently be thought of as paid to the government in units of domestic currency or in units of foreign currency. In the former case, households must acquire on foreign exchange markets the domestic currency needed to pay the tax; in the latter case, the government acquires domestic currency by selling the foreign currency. In either case, the lump-sum refund of tax reve-

<sup>7</sup>The second equation in (10) would be

$$N_t(s) = n_{t-1}(s) + [q_{t-1}(s)\bar{y}_{t-1}k + F_t(s)],$$

if the term in brackets were negative, to prevent subsidies on sales of foreign currencies. We will restrict our attention to the case in which the term in brackets is nonnegative, which rules out some probability distributions on the state vector that might result in sales of foreign currencies in some states.



nue to households is paid in domestic currency.

Arbitrage implies that the exchange rate in state  $s$  of date  $t$  given in equation (8) equals  $P_{Ft}(s)/P_{Bt}(s)$ , which is the relative price at which households can contract, at date zero, to exchange foreign and domestic moneys in the future on a state-contingent basis. Because markets are complete, future trade in assets is redundant. This argument also applies to the government. Because the government can contract at date zero to exchange currencies in the future on a state-contingent basis, spot market currency trades are also redundant for government transactions.

Substituting equations (10) and (9) into (2) to eliminate  $x_t(s)$  and  $y_t(s)$ , the domestic household chooses  $M_0$ ,  $N_0$ ,  $H$ ,  $K$ ,  $B_t(s)$ ,  $F_t(s)$ ,  $m_t(s)$ , and  $n_t(s)$  to maximize expression (2) subject to equations (5) and (9). Necessary conditions for the domestic household are, in addition to the constraints:

$$(12) \quad U_1(x_0, y_0) = p_0 \lambda P_{M0}$$

$$(13) \quad U_2(x_0, y_0) = q_0 \lambda P_{N0}(1 + \tau_0)$$

$$(14) \quad \sum_{t=1}^{\infty} \beta^t E[U_1(x_t(s), y_t(s)) \times p_{t-1}(s) \bar{x}_{t-1}/p_t(s)] = \lambda P_{H0}$$

$$(15) \quad \sum_{t=1}^{\infty} \beta^t E[U_2(x_t(s), y_t(s)) \times q_{t-1}(s) \bar{y}_{t-1}/q_t(s)(1 + \tau_t)] = \lambda P_{K0}$$

$$(16) \quad \beta^t U_1(x_t(s), y_t(s)) = p_t(s) \lambda P_{Bt}(s)$$

$$(17) \quad \beta^t U_2(x_t(s), y_t(s)) = q_t(s)(1 + \tau_t) \lambda P_{Ft}(s)$$

$$(18) \quad \mu_t(s) = U_1(x_t(s), y_t(s))/p_t(s) - \beta E[U_1(x_{t+1}(s), y_{t+1}(s))/p_{t+1}(s)]$$

$$(19) \quad \nu_t(s) = U_2(x_t(s), y_t(s))/q_t(s) - \beta E[U_2(x_{t+1}(s), y_{t+1}(s)q_{t+1}(s))]$$

$$(20) \quad m_t(s)\mu_t(s) = 0, \mu_t(s) \geq 0$$

$$(21) \quad n_t(s)\nu_t(s) = 0, \nu_t(s) \geq 0,$$

where  $\lambda$  is the multiplier on the budget constraint (5). Equations (16)–(21) hold for all  $t=1, 2, \dots$ . The redundancy of trade in shares of firms follows directly from equations (14)–(17). Summing over states and time periods we find that

$$(22) \quad P_{H0} = \sum_{t=1}^{\infty} \int p_{t-1}(s) \bar{x}_{t-1} P_{Bt}(s) \pi(ds).$$

$$P_{F0} = \sum_{t=1}^{\infty} \int q_{t-1}(s) \bar{y}_{t-1} P_{Ft}(s) \pi(ds).$$

These equations can be updated to price equities in any time period; they imply that any trade in equities can be exactly duplicated with trade in the other contingent assets.

From equation (3) and the foreign analogues to equations (5)–(11), we can define the foreign household optimization problem and obtain an analogous set of first-order conditions among which are (with the obvious notation)

$$(23) \quad \beta^t U_1^*(x_t^*(s), y_t^*(s)) = \lambda^* p_A(s)(1 + \tau_t^*) P_{BA}(s),$$

$$\beta^t U_2^*(x_t^*(s), y_t^*(s)) = \lambda^* q_A(s) P_{FA}(s).$$

## II. Equilibrium

Equilibrium requires that world demands and supplies of  $X$  and  $Y$  are equated in each state in each period. Equilibrium conditions, along with equations (12)–(21), their foreign counterparts, and an arbitrary choice of numeraire for the prices, determine all prices, consumptions, and productions as functions of these Lagrange multipliers. The multipliers, in turn, are determined through the budget constraints and transversality conditions, and are functions of the distribution of wealth between the two countries. It is convenient to choose a normalization so that

the domestic multiplier  $\lambda$  is unity.<sup>8</sup> Then, loosely speaking, the domestic country is wealthier or less wealthy than the foreign country as  $\lambda^*$  is larger or smaller than one.

Combining equations (16), (17), (23), and equilibrium conditions for product markets, we obtain

$$\begin{aligned} (24) \quad U_1^*(\bar{x}_t - x_t(s), \bar{y}_t - y_t(s)) \\ = \lambda^*(1 + \tau_t^*)U_1(x_t(s), y_t(s)) \\ (1 + \tau_t)U_2^*(\bar{x}_t - x_t(s), \bar{y}_t - y_t(s)) \\ = \lambda^*U_2(x_t(s), y_t(s)). \end{aligned}$$

Note that only current taxes on foreign-currency acquisitions affect current consumptions. This results from an intertemporally separable utility function and the absence of real investment in the model. In a more general model that relaxed these features, the conditional probability distribution of future taxes, given current taxes, would also affect current allocations and trade.

Define  $T_t = 1 + \tau_t$  and  $T_t^* = 1 + \tau_t^*$ . Total differentiation of equations (24)—holding fixed  $\lambda^*$  for a comparison across states—gives

$$\begin{aligned} (25) \quad \begin{pmatrix} dx_t \\ dy_t \end{pmatrix} &= \frac{1}{\Delta} \\ &\times \begin{bmatrix} -U_2^*(U_{12}^* + \lambda^*T_t^*U_{12}) & -\lambda^*U_1(T_tU_{22}^* + \lambda^*U_{22}) \\ U_2^*(U_{11}^* + \lambda^*T_t^*U_{11}) & \lambda^*U_1(T_tU_{12}^* + \lambda^*U_{12}) \end{bmatrix} \\ &\times \begin{pmatrix} d\tau_t \\ d\tau_t^* \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} (26) \quad \Delta &= T_t[U_{11}^*U_{22}^* - U_{12}^{*2}] \\ &+ \lambda^{*2}T_t^*[U_{11}U_{22} - U_{12}^2] \\ &+ \lambda^*[U_{11}^*U_{22} - U_{12}^*U_{12}] \\ &+ \lambda^*T_tT_t^*[U_{11}U_{22}^* - U_{12}U_{12}^*], \end{aligned}$$

<sup>8</sup>This normalization is permitted because all prices in equation (5) were expressed in terms of an arbitrary numeraire.

and  $\Delta > 0$  by the assumption in expression (4) and quasiconcavity of the utility functions.

We now investigate how alternative realizations of the exogenous random variables affect equilibrium allocations and prices. In particular, we are interested in the differences between variables such as  $x_t(s)$  and  $x_t(s')$ , where  $s$  and  $s'$  are identical sequences except that  $\tau_t \neq \tau_t'$ .

Equation (25) implies that<sup>9</sup>

$$\begin{aligned} (27) \quad \text{sign} \frac{\partial x_t(s)}{\partial \tau_t} &= -\text{sign} \frac{\partial x_t^*(s)}{\partial \tau_t} \\ &= -\text{sign}[U_{12}^* + \lambda^*T_t^*U_{12}] \\ \frac{\partial y_t(s)}{\partial \tau_t} &< 0, \quad \frac{\partial y_t^*(s)}{\partial \tau_t} > 0 \end{aligned}$$

and

$$\begin{aligned} (28) \quad \frac{\partial x_t^*(s)}{\partial \tau_t^*} &> 0, \quad \frac{\partial x_t(s)}{\partial \tau_t^*} < 0 \\ \text{sign} \frac{\partial y_t(s)}{\partial \tau_t^*} &= -\text{sign} \frac{\partial y_t^*(s)}{\partial \tau_t^*} \\ &= \text{sign}[U_{12}^* + \lambda^*T_t^*U_{12}]. \end{aligned}$$

The results in equations (27) and (28) show that states and time periods with greater domestic taxes on income from foreign as-

<sup>9</sup>Our results in equation (25) can be used to calculate the approximate covariances of consumption and taxes implied by the model for any arbitrary (stationary) probability distribution on taxes, along the lines of Svensson, 1985b, and Stockman and Svensson, 1987. For example, the covariance of  $x_t$  and  $\tau_t$  is approximately

$$\begin{aligned} \text{COV}(x_t(s), \tau_t) &= -\frac{U_2^*}{\Delta}[U_{12}^* + \lambda^*T_t^*U_{12}]\sigma_{\tau\tau}^* \\ &- \frac{\lambda^*U_1}{\Delta}[TU_{22}^* + \lambda^*U_{22}]\sigma_{\tau\tau}^*, \end{aligned}$$

where  $\sigma_{\tau\tau}^2$  is the variance of  $\tau$  and  $\sigma_{\tau\tau}^*$  is the covariance of  $\tau$  and  $\tau^*$ .

sets and purchases of foreign currency are associated with lower domestic consumption of foreign goods and lower, unchanged, or greater consumption of domestic goods as a weighted sum of  $U_{12}$  and  $U_{12}^*$  is greater than, equal to, or less than zero. This contrasts with the more common argument that taxation of foreign-currency acquisitions will have some expenditure-switching effects that will increase consumption of exportables irrespective of the sign of  $U_{12}$ ; that effect is present here only if  $U_{12}$  is negative. Furthermore, *ex post* utility in period  $t$  is generally decreased by these taxes on acquisitions of foreign currency, since

$$(29) \quad \frac{dU}{d\tau_t} = \frac{U_2^*}{\Delta} \{ U_2 [U_{11}^* + \lambda^* T_t^* U_{11}] - U_1 [U_{12}^* + \lambda^* T_t^* U_{12}] \}$$

is negative unless  $U_{12}^* + \lambda^* T^* U_{12}$  is negative and very large.

These results differ substantially from those obtained in models without contingent assets or in which households are assumed to ignore the possibility of changes in government policies. In either of these cases, an increase in the domestic tax rate on foreign-currency acquisitions not only leads domestic households to substitute away from foreign currency into domestic currency and other assets, but affects the distribution of wealth. Substitution out of foreign currency reduces the demand for the foreign good and lowers its relative price. This redistributes wealth from owners of foreign firms to owners of domestic firms. If equities are traded internationally, the effects on domestic and foreign wealth depend on portfolio shares. If it is assumed that equities are all held domestically (for example, because they cannot be traded internationally), then the domestic tax increase raises domestic wealth and lowers foreign wealth. Therefore, the tax increase raises domestic consumption of  $X$  even with separable utility. If the tax is small enough, it also raises domestic utility (through an argument somewhat like that for an optimal tariff). In our model, in contrast, domestic consumption of  $X$  is unaffected if

utility is separable, and more generally could rise or fall (depending on  $U_{12}$ ). Domestic *ex post* utility falls, except in the special case noted below equation (29), because financial markets eliminate the wealth redistribution.

We can also calculate the effect on *ex post* foreign utility at time  $t$  of a change in the domestic tax rate. From equation (25)

$$(30) \quad \frac{dU^*}{d\tau_t} = -\frac{U_2^*}{\Delta} \{ U_2^* [U_{11}^* + \lambda^* T_t^* U_{11}] - U_1^* [U_{12}^* + \lambda^* T_t^* U_{12}] \},$$

which is positive unless  $U_{12}^* + \lambda^* T_t^* U_{12}$  is negative and very large.

The competitive equilibrium model described in this paper is equivalent to a social welfare maximization model in which social welfare is defined as a weighted average of the two representative households' utilities,

$$(31) \quad W(\bar{x}, \bar{y}) = \max U(x, y) + \lambda^{*-1} U^*(x^*, y^*)$$

subject to  $(x, y) + (x^*, y^*) = (\bar{x}, \bar{y})$ . Taking this as our measure of world welfare, equations (24), (29), and (30) imply

$$(32) \quad \frac{\partial W}{\partial \tau_t} = \frac{U_2^*}{\Delta} \left\{ [U_{11}^* + \lambda^* T^* U_{11}] \times U_2 \left( 1 - \frac{\lambda^*}{T} \right) - [U_{12}^* + \lambda^* T^* U_{12}] \times U_1 (1 - \lambda^* T^*) \right\},$$

which can take either sign. If  $\lambda^* = 1$ ,  $U_{12} \leq 0$ ,  $U_{12}^* \leq 0$ , and domestic and foreign tax rates are initially positive, then the expression in equation (32) is negative. More generally, the sign depends on the initial distribution of wealth across countries and on  $U_{12}$  and  $U_{12}^*$ . Even if utility functions are separable, the direction of the effect on world welfare of an increase in the tax rate depends on the wealth distribution through the  $\lambda^*$  term.

The terms of trade can be derived as usual from the first-order and equilibrium conditions

$$\begin{aligned}
 (33) \quad Q_t(s) &\equiv \frac{p_t(s)}{e_t(s)q_t(s)} \\
 &= \frac{U_1(x_t(s), y_t(s))}{U_2(x_t(s), y_t(s))} (1 + \tau_t) \\
 &= \frac{U_1^*(\bar{x}_t - x_t(s), \bar{y}_t - y_t(s))}{U_2^*(\bar{x}_t - x_t(s), \bar{y}_t - y_t(s))} \\
 &\quad \times (1 + \tau_t^*)^{-1}.
 \end{aligned}$$

The last expression in equation (33) is useful for deriving the effect of an increase in the domestic tax rate on the terms of trade. We obtain

$$\begin{aligned}
 (34) \quad \frac{\partial Q_t}{\partial \tau_t} &= \frac{Q\lambda^*}{\Delta} \{ (U_{11}^*U_{22}^* - U_{12}^{*2})/\lambda^* \\
 &\quad + (U_{11}U_{22}^*T_t - U_{12}U_{12}^*T_t^*) \\
 &\quad + (U_{11}^*U_{12} - U_{11}U_{12}^*T_t/T_t^*)/Q \}.
 \end{aligned}$$

The first two terms in equation (34) reflect the effect of a change in the domestic tax on  $U_2^*$ . The first term is positive and the second term is positive unless inequality (4) is nearly violated and  $T^*/T$  is sufficiently large. Because an increase in the domestic tax rate unambiguously raises  $y^*$ , it tends to reduce  $U_2^*$  and raise the domestic terms of trade. The third term reflects the effect of a change in  $\tau$  on  $U_1^*$ . Because an increase in the domestic tax rate has an ambiguous effect on  $x_t^*$ , as shown in equation (27), the effect on  $U_1^*$  and the terms of trade are also ambiguous. If the utility functions were separable, the third term would vanish and an increase in domestic taxes would unambiguously raise the domestic terms of trade. However, the third term could take a negative sign if either  $U_{12}$  is large and positive or  $U_{12}^*$  is large and negative. In the former case, the fall in domestic consumption of  $Y$  reduces domestic marginal utility and consumption of  $X$ ,

and this is absorbed in equilibrium by a rise in foreign consumption of  $X$ . In the latter case, the increase in foreign consumption of  $Y$  reduces foreign consumption of  $X$  directly by lowering its marginal utility. Even if the third term is negative, however, equation (34) is likely to be positive so that an increase in domestic taxes on acquisitions of foreign currency raises the terms of trade.

The effects of exchange controls and capital controls on nominal prices and the exchange rate can be determined from the other necessary conditions and equilibrium conditions. Notice that the equilibrium allocations derived in Section I are independent of the behavior of nominal prices. Given these allocations, and given  $\lambda^*$ , which depends upon the international distribution of wealth at date zero, we have a system of equations consisting of equations (18), (19), (20), and (21) for every date  $t$  and state  $s$ , in the variables  $m_t(s)$ ,  $n_t(s)$ , the multipliers  $\mu$  and  $\nu$ , and the prices  $p_t(s)$  and  $q_t(s)$ . We have an exactly analogous set of equations from the necessary conditions for the foreign optimization problem; the endogenous variables in those equations are  $m_t^*(s)$ ,  $n_t^*(s)$ , the foreign multipliers  $\mu^*$  and  $\nu^*$ , and the prices  $p_t^*(s)$  and  $q_t^*(s)$ . Finally, we have the equilibrium conditions for each money, which can be written as

$$\begin{aligned}
 (35) \quad p_t(s) &= [\bar{M} - m_t(s) - m_t^*(s)]/\bar{x}_t \\
 q_t(s) &= [\bar{N} - n_t(s) - n_t^*(s)]/\bar{y}_t,
 \end{aligned}$$

where  $\bar{M}$ ,  $\bar{N}$  are the fixed per capita money supplies.

The exchange rate can be derived from equations (33) and (35):

$$\begin{aligned}
 (36) \quad e_t(s) &= \frac{[\bar{M} - m_t(s) - m_t^*(s)]}{[\bar{N} - n_t(s) - n_t^*(s)]} \\
 &\quad \times \frac{\bar{y}_t}{\bar{x}_t} \frac{U_2^*(x_t^*(s), y_t^*(s))}{U_1^*(x_t^*(s), y_t^*(s))} (1 + \tau_t^*).
 \end{aligned}$$

Notice that this solution for the exchange rate is analogous, aside from the tax distortion term, to the expression derived by Lucas

(1982). The new term involves a tax wedge associated with exchange controls and capital controls. Variations in this tax term could, according to our theory, play an important role in exchange rate variations.

In Lucas's paper there is no precautionary demand for money because nominal interest rates are assumed to be strictly positive and, as in our model, all relevant information about the subsequent product market is available during asset market trade. The absence of precautionary demands for moneys implies that velocities of money are unity. In our model, after-tax interest rates differ across countries. Domestic households do not pay taxes on income from domestic assets (they make payments in domestic currency) but foreign households do. The domestic one-period nominal interest rate at time  $t$  is, using equation (16).

$$(37) \quad 1 + i_t(s) = \frac{U_1(x_t(s), y_t(s))/p_t(s)}{\beta E_t[U_1(x_{t+1}(s), y_{t+1}(s))/p_{t+1}(s)]}$$

This result, with equation (18) implies that the multiplier  $\mu_t(s)$  is strictly positive if and only if the domestic nominal interest rate is positive,

$$(38) \quad \mu_t(s) = [i_t(s)/(1 + i_t(s))] \times U_1(x_t(s), y_t(s))/p_t(s).$$

As in Lucas (1982), we could restrict parameter values in such a way that  $i_t(s)$  (and so  $\mu_t(s)$ ) is strictly positive.<sup>10</sup> However,

<sup>10</sup>Although the velocity of money is endogenously variable in our model, the choice of money holdings of each date is made with full knowledge of the current state vector. The effects of exchange and capital controls on real allocations in our model do not, therefore, affect allocations only by including households to choose money holdings that, *ex post*, are too large. In fact, even if we chose probability distributions such that velocity were fixed at unity, the effects on consumption in equations (27) and (28) would still be obtained. We could follow Svensson (1985a,b) and restrict some information about future product markets during asset

in our model that restriction does not imply a unit of velocity of money because foreign households face potential future taxes on acquisitions of domestic money, including interest income on domestic assets. If foreign households face current taxes that are low in comparison to expected future taxes (loosely speaking), they will acquire domestic money currently and hold it for future use, which offsets the velocity of domestic money and domestic nominal prices. Similarly, domestic households may acquire foreign currency to hold if current domestic taxes are low relative to those anticipated in the future. The foreign analogue to equation (18) is

$$(39) \quad \mu_t^*(s) = U_1^*(x_t^*(s), y_t^*(s))/p_t(s) - \beta E_t[U_1^*(x_{t+1}^*(s), y_{t+1}^*(s))/p_{t+1}(s)] = [\hat{i}_t(s)/(1 + \hat{i}_t(s))] \times U_1^*(x_t^*(s), y_t^*(s))/p_t(s),$$

where the after-tax domestic nominal interest rate to foreign households,  $\hat{i}_t$ , is given by

$$(40) \quad 1 + \hat{i}_t(s) = \frac{U_1^*(x_t^*(s), y_t^*(s))/p_t(s)}{\beta E_t[U_1^*(x_{t+1}^*(s), y_{t+1}^*(s))/p_{t+1}(s)]}$$

A strictly positive domestic nominal interest rate does not guarantee that  $\hat{i}_t(s)$  and  $\mu_t^*(s)$  cannot be zero. Using equations (16) and

market trade. This would result in strictly positive nominal interest rates along with variable velocity of money. If our model is changed so that endowments at date  $t$  are realized only *after* asset market trade of date  $t$ , then the resulting equations analogous to (24) are

$$(24') \quad E_t[U_1^*(\bar{x}_t - x_t(s), \bar{y}_t - y_t(s))/p_t(s)] = \lambda^*(1 + \tau_t^*) E_t[U_1(x_t(s), y_t(s))/p_t(s)], \\ (1 + \tau_t) E_t[U_2^*(\bar{x}_t - x_t(s), \bar{y}_t - y_t(s))/q_t(s)] = \lambda^* E_t[U_2(x_t(s), y_t(s))/q_t(s)].$$

(23) we can write

$$(41) \quad \frac{1 + i_t(s)}{1 + \hat{i}_t(s)} = \frac{\int P_{Bt+1}(s)(1 + \tau_{t+1}^*)\pi(ds)}{(1 + \tau_t^*)\int P_{Bt+1}(s)\pi(ds)}$$

so  $i > \hat{i}$  when the current foreign tax rate is smaller than a weighted average of future possible tax rates. In this situation,  $m_t^*(s)$  will be positive and, according to equation (35), the domestic nominal price level will be lower than otherwise.

The result that speculation about future changes in tax rates affects money demands and nominal prices contrasts sharply with the result in equation (24) regarding real allocations. The solutions for consumption and relative prices depend only on the current level of taxes and, through  $\lambda^*$ , the unconditional probability distribution of exogenous disturbances. The solutions for money demands and nominal prices depend also on the conditional probability distribution of disturbances, given the current state.

The effects of exchange controls and capital controls on the exchange rate, given in equation (36), can be broken into two main effects. First, they affect nominal price levels; this is the contribution of the first term in equation (36). Second, they affect the exchange rate by altering the terms of trade as in equation (34). The net effect on the exchange rate is ambiguous. If  $i_t(s)$ ,  $\hat{i}_t(s)$ , and the analogous foreign interest rates  $i_t^*(s)$  and  $\hat{i}_t^*(s)$  are all positive, then domestic currency appreciates or depreciates with a rise in  $\tau_t$  as the terms of trade rise or fall, that is, as expression (34) is positive or negative. The discussion following equation (34) implies that domestic currency is likely to appreciate with greater exchange controls, though, as previously shown, utility falls.

### III. Conclusions

International financial markets have become increasingly sophisticated, but they remain subject to taxes and government con-

trols. This raises the question of how changes in taxes on financial markets—either toward stricter controls or toward more liberalization—affect portfolios, prices, consumption, and international trade.

We have found that when sophisticated financial markets are available, changes in taxes on acquisitions of foreign currency affect trade and prices differently than in the absence of those financial markets. An increase in taxes on foreign currency acquisitions raises the gross cost of importing foreign goods by taxing the foreign money required to purchase them. This reduces imports and raises the terms of trade. Financial markets allow households to shift wealth across states of the world, so the wealth effect of the increase in terms of trade is nil if financial markets are sufficiently developed. As a consequence, although imports fall, domestic demand is not shifted to exportables unless the two goods are direct substitutes in utility; if utility is separable, exports are unaffected. The rise in the terms of trade requires an appreciation of domestic currency unless nominal export prices are affected. An increase in the current level of domestic taxes lowers the speculative demand for foreign currency if that increase is expected to be temporary; this raises the foreign price level and reinforces the appreciation of domestic currency. Similar results hold if higher domestic taxes raise the conditional probability that the foreign country will respond in the future by imposing its own exchange controls: the foreign speculative demand for domestic currency increases immediately in order to avoid the prospect of future controls, and this lowers the domestic price level and reinforces the domestic currency appreciation.

Our results were obtained from a model with an unrealistically large set of contingent claims markets. Private information or other reasons for additional restrictions on these markets were neglected in the interest of simplicity. A more realistic model would include restrictions generated by moral hazard, adverse selection, and time-consistency problems. The conclusions, however, are likely to be similar to these derived above as long as

there remain opportunities to trade in assets whose real returns are affected by policy actions. The development of international financial markets in recent years adds to the likely relevance of our results. Many real-world assets, including stocks, nominal bonds and Eurodeposits, forward contracts, future contracts, currency options, futures options, currency swaps, interest-rate swaps, etc., have real returns that depend on government policies such as exchange and capital controls. Corporations use swaps routinely, and can thereby create almost any contingent asset desired. Equities in firms that trade internationally or that utilize swaps or the other assets mentioned above provide indirect access to these markets for households who may not participate in them directly. Because these assets permit trades that are contingent on events that include exchange and capital controls, the effects of changes in controls are likely to be similar to those discussed above.

Our model also suggests a new direction of research in the theory of exchange rate behavior. Changes in tax rates on income from financial assets, and analogous regulations and controls can have major effects on exchange rates.

## REFERENCES

- Adams, C. and Greenwood, J., "Dual Exchange Rate Systems and Capital Controls: An Investigation," *Journal of International Economics*, February 1985, 18, 43-63.
- Cooley, T., LeRoy, S. and Raymon, N., (1984a) "Econometric Policy Evaluation: Note," *American Economic Review*, June 1984, 74, 467-70.
- \_\_\_\_\_, (1984b) "Modeling Policy Interventions," unpublished paper, University of California-Santa Barbara.
- Cumby, R., "Monetary Policy under Dual Exchange Rates," NBER Working Paper No. 1424, 1984.
- Greenwood, J. and Kimbrough, K., "Capital Controls and International Transmission of Fiscal Policy," *Canadian Journal of Economics*, November 1985, 18, 743-65.
- \_\_\_\_\_, "An Investigation in the Theory of Foreign Exchange Controls," *Canadian Journal of Economics*, May 1987, 20, 271-88.
- Helpman, E., "An Exploration in the Theory of Exchange Rate Regimes," *Journal of Political Economy*, October 1981, 89, 865-90.
- \_\_\_\_\_, and Razin, A., "The Role of Saving and Investment in Exchange Rate Determination under Alternative Monetary Mechanisms," *Journal of Monetary Economics*, May 1984, 13, 307-26.
- \_\_\_\_\_, "Dynamics of a Floating Exchange Rate Regime," *Journal of Political Economy*, August 1982, 90, 728-54.
- Lucas, R., "Econometric Policy Evaluation: A Critique," in K. Brunner and A. H. Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, *Carnegie-Rochester Series on Public Policy*, *Journal of Monetary Economics*, Supplementary Series 1976, 1, 19-46.
- \_\_\_\_\_, "Interest Rates and Currency Prices in a Two-Country World," *Journal of Monetary Economics*, November 1982, 10, 335-60.
- \_\_\_\_\_, and Stokey, N., "Optimal Fiscal Policy in an Economy without Capital," *Journal of Monetary Economics*, July 1983, 12, 55-93.
- Obstfeld, M., "Capital Controls, the Dual Exchange Rate, and Devaluation," NBER Working Paper No. 1324, 1984.
- Sims, C., "Policy Analysis with Econometric Models," *Brookings Papers on Economic Activity*, 1: 1982, 107-64.
- \_\_\_\_\_, "A Rational Expectations Framework for Short-Run Policy Analysis," unpublished paper, University of Minnesota, 1985.
- Stockman, A., "A Theory of Exchange Rate Determination," *Journal of Political Economy*, August 1980, 88, 4, 673-98.
- \_\_\_\_\_, "Fiscal Policy and International Financial Markets," forthcoming in Jacob Frenkel, ed., *International Aspects of Fiscal Policy*, Chicago: University of Chicago Press, 1987.
- \_\_\_\_\_, and Dellas, H., "Asset Markets, Tariffs, and Political Risk," *Journal of International Economics*, November 1986, 21,

- 199-214.
- \_\_\_\_\_ and Svensson, L., "Capital Flows, Investment, and Exchange Rates," *Journal of Monetary Economics*, March 1987, 19, 171-202.
- Svensson, L., (1985a) "Currency Prices, Terms of Trade, and Interest Rates: A General Equilibrium Asset-Pricing, Cash-in-Advance Approach," *Journal of International Economics*, February 1985, 18, 17-41.
- \_\_\_\_\_, (1985b) "Money and Asset Pricing in a Cash-in-Advance Economy," *Journal of Political Economy*, October 1985, 93, 919-44.



# Trade in Risky Assets

By LARS E. O. SVENSSON\*

*A theory of the international trade pattern in risky assets is developed by applying the law of comparative advantage to asset trade. According to this law there is a tendency for a country to import assets that have relatively high autarky prices. It is examined how autarky prices are affected by international differences in (i) stochastic properties of output/endowments, (ii) the rate of time preference, (iii) the degree of risk aversion, and (iv) subjective beliefs, and how such differences predict overall capital account deficits or surpluses as well as the composition of the capital account into trade in specific risky assets.*

This paper develops a simple but general theory of the determinants of the international pattern of trade in risky assets. The importance of international trade in risky assets is obvious, with increased liberalization of international capital movements, and with the observation that in practice all assets are risky in the sense that their real returns are uncertain.<sup>1</sup> Yet it seems that there is much less research done on the pattern of trade in explicitly risky assets than on the pattern of trade in goods.

The theory is developed by borrowing from and synthesizing several strands of literature. We start from the modern formulations of standard international trade theory, more precisely the general law of comparative advantage as developed by Alan Deardorff

(1980) and Avinash Dixit and Victor Norman (1980). According to the law of comparative advantage there is a positive correlation between a country's net import of goods and the country's autarky prices relative to world prices (or relative to autarky prices in the rest of the world), such that on average a country is a net importer of goods for which autarky prices are relatively high. With only two goods, the law of comparative advantage provides an exact relation between the trade pattern and relative autarky prices. With more than two goods, it provides only a correlation between the vectors of net import and relative autarky prices, and it does not provide an exact relation for each individual good.

It is well known that the standard trade theory can be extended to an intertemporal theory of international borrowing and lending, by interpreting commodities as dated goods. The law of comparative advantage then implies that a country will on average have a trade surplus in periods for which the autarky present value of goods is relatively high, that is, for which autarky interest rates are relatively low.<sup>2</sup> It is also clear that the standard trade theory can be extended to the case with uncertainty, where goods are distinguished by the state of the world in which they occur.<sup>3</sup> The principle of comparative

\*Institute for International Economics, University of Stockholm, S-106 91 Stockholm, Sweden. I am very grateful for comments by Alan Deardorff, Avinash Dixit, Jeremy Greenwood, Boyan Jovanovic, Peter Howitt, Hal Varian, Sweder van Wijnbergen, two anonymous referees (in particular one with an extremely thorough and constructive report), and participants in seminars at University of Western Ontario, University of Michigan, World Bank, NBER Summer Institute, and IIES. Remaining errors and obscurities are my own responsibility.

<sup>1</sup>Stocks and equities are obviously risky assets, but so are all nominal bonds in any currency since there is exchange rate and price-level risk. Exchange rate risk makes even very short-term bank deposits risky. A nonrisky asset would be a hypothetical<sup>2</sup> appropriately indexed (to some consumer price index, say) short-term deposit. Even such an asset is not sure in utility terms (see fn. 16).

<sup>2</sup>For an explicit statement of the intertemporal extension of the standard trade theory, see Torsten Persson and Alan Stockman, 1987.

<sup>3</sup>See John Pomery, 1984; Elhanan Helpman, 1985a; and Persson and Stockman, 1987.

advantage then says that a country will on average import goods in states for which the autarky prices for Arrow-Debreu securities, that is, state-contingent deliveries, are relatively high.

A special case of trade in risky assets has received considerable interest. This is trade in claims to firms' profits, equity. After pioneering work by Elhanan Helpman and Assaf Razin (1978), a number of papers have recently examined the effects on trade in equities on welfare, resource allocation, and the goods trade pattern.<sup>4</sup>

Here we will reformulate the law of comparative advantage so as to cover the case of trade in any arbitrarily specified set of assets, complete or incomplete.<sup>5</sup> This will allow us to include as special cases trade in sure indexed bonds, trade in Arrow-Debreu securities, and trade in equities, or rather claims to firms' output (we shall make a simplifying assumption of exogenous stochastic outputs/endowments and no inputs, so as to be able to disregard the effect of trade in assets on production decisions, in which case claims to profit and claims to output coincide).

In standard trade theory, there are basically two approaches to examine the determinants of the trade pattern. One, the comparative advantage approach, is to start from the law of comparative advantage and its emphasis on autarky price differences, and then to go behind the autarky price differences and explain how these are caused by underlying differences between countries with respect to technology, endowments, preferences, or other characteristics. The other, the "direct" approach, is to look di-

rectly at trade equilibria without any reference to autarky prices, and infer how differences between countries directly determine the trade pattern. Whereas the autarky prices approach was common in the early work on the goods trade pattern,<sup>6</sup> the direct approach has more recently been the dominant one, both in standard trade theory and in the literature on trade in equities referred to above.<sup>7</sup>

There is, however, a special reason for basing a theory of the trade pattern for risky assets on relative autarky prices. The reason is that we can borrow from the general-equilibrium asset-pricing theory developed by Robert Lucas, 1978, Douglas Breeden, 1979, and others. It turns out to be very convenient to use this theory in order to express autarky asset prices in terms of autarky real interest rates and risk premia. Our work is hence closely related to international applications of this asset pricing theory, for instance, Lucas, 1982; René Stulz, 1981, 1984; Svensson, 1985; and Stockman and Svensson, 1987. That literature has focused on the determinants of prices on internationally traded risky assets, but not examined the trade pattern in risky assets in itself. In the typical set up, as in Lucas, 1982, there is trade in the outside assets, namely claims to output (equities), currencies, and claims to government transfers. Since representative consumers with identical preferences are assumed, there is no trade in other, inside assets (which does not prevent any arbitrary inside asset to be priced, however). Furthermore, the trade pattern in the existing outside assets is trivial, since a perfectly pooled equilibrium is assumed, in which all investors hold the same portfolio.<sup>8</sup> In our analy-

<sup>4</sup>See, for instance, Pomery, 1984, and later work by Helpman, 1985a,b, and Gene Grossman and Razin, 1984, 1985. Harold Cole, 1986, examines the effect of trade in different kinds of assets (*ex post* securities, Arrow-Debreu state-contingent deliveries, and Helpman-Razin equities) on variance and covariance of key real variables, like output, consumption, and trade balance.

<sup>5</sup>The set of assets is complete (incomplete) in the usual sense of having at least as many (fewer) linearly independent assets as (than) the number of states of the world.

<sup>6</sup>See, for instance, the classic paper by Ronald Jones, 1956.

<sup>7</sup>For examples of use of the direct approach to the determinants of the pattern of trade, see Alan Deardorff, 1982; the survey by Wilfred Ethier, 1984; Dixit and Alan Woodland, 1982; and James Markusen and Lars Svensson, 1985, for trade in goods, and my paper, 1984; and Ethier and Svensson, 1986, for trade in goods and factors.

<sup>8</sup>That is, relative to autarky each country (in a two-country world) exports half of its assets and im-

sis, equilibria will generally not be perfectly pooled.

We mentioned that our theory is general in the sense of covering any arbitrary complete or incomplete set of assets, including as special cases sure indexed bonds, equities, and claims on output (stocks), and Arrow-Debreu securities. Also, our theory includes the determinants of the aggregate current account and capital account, hence aggregate international borrowing and lending, as well as the composition of the capital account, the trade in individual assets (subject to the qualification that when there are many assets, results are in the form of correlations and hold on average, but not exactly for each individual asset).

The first step in our method is to express the autarky asset price for a given asset in terms of the autarky real interest rate and the autarky risk measure (the risk measure is the product of the risk premium and the asset price). Differences in countries' autarky real interest rates affect the autarky prices of and trade in all assets, and are related to whether a country has an overall capital account deficit or surplus and hence is a net

lender or borrower. A country with a relatively low autarky real interest rate has a tendency to have an overall capital account deficit and be a net lender. Differences in autarky risk measures are specific to individual assets and are related to the trade in individual assets. A country with a relatively low autarky risk measure for an asset (that is, for which an asset is relatively less risky) has a tendency to import that asset.

The second step is to examine what determines the differences between countries' autarky real interest rates and risk measures. We will look at the effect on autarky real interest rates and risk measures of differences between countries with respect to technology, endowments, and preferences; or more precisely (i) the stochastic properties of output/endowments, (ii) the rate of time preference, (iii) the degree of risk aversion, and (iv) expectations (subjective probability beliefs).

The paper is organized as follows. Sections I and II deal with preliminaries and can be skimmed by readers not interested in the standard derivation of the law of comparative advantage. Section I describes the model; the equilibrium for a single country, and demonstrates gains from trade in risky assets. Section II describes a world equilibrium with two countries and derives the law of comparative advantage for trade in risky assets. Section III, the core of the paper, discusses the determination of autarky asset prices, derives the effect of cross-country differences in technology/endowments and preferences on autarky real interest rates and risk measures, and finds the trade pattern for arbitrary assets as well as the special cases of sure bonds, stocks, and Arrow-Debreu securities. Section IV concludes.

The results are summarized in a highlighted paragraph at the end of each subsection of Section III. Reading just those paragraphs gives an overview of the results.

### I. Equilibrium in a Single Country and Gains from Asset Trade

We consider a situation with one good and two periods. There are two countries, home and foreign, in the world. Period 1

ports half of the other country's assets. Still, capital movements, and correlations between key macro variables like investment, the current account, output, etc., can be studied, as in Stockman and Svensson, 1987, but any current and capital account movements are due exclusively to *revaluation* of domestically based assets relative to foreign based assets, not to changes in the ownership of assets.

Bernard Dumas (1986) considers a model with two investors with different degrees of risk aversion where the investors' portfolios are revised over time and asset trades between them occur. Stockman and Harris Dellas, 1986, and Stulz, 1986, consider international asset pricing models with nontraded goods, where consumers do not have perfectly pooled equilibria but hold a larger share of domestic assets. Their focus is exclusively on equilibrium asset price and exchange rate determination and variability. Stockman and Alejandro Hernández D., 1986, utilize an international asset pricing model to demonstrate that the effect on policy like capital controls depends crucially on whether the private sector can hedge against the policy by trading in risky assets (in their case Arrow-Debreu securities). Robert Gordon and Hal Varian, 1986, discuss welfare effects of taxes on internationally traded risky assets in a CAPM model and examine the analogue to the optimum tariff result for trade in goods.

outputs in the home and foreign country,  $y^1$  and  $y^{*1}$ , are exogenous, and deterministic. Period 2 outputs in the two countries,  $y^2$  and  $y^{*2}$ , are also exogenous, but stochastic. We call the vector  $s = (y^2, y^{*2})$  the state of the world in period 2. Goods are perishable and there is no storage or other investment technology.

There is a given set  $J$  of  $J$  different assets. (We let  $J$  denote both the set and the number of elements of the set.) These assets are traded on a world asset market in period 1, before the uncertainty about the state of the world in period 2 is resolved. Each asset  $j \in J$  is characterized by a given (gross real) return function  $R_j(s)$ , which expresses the gross real returns paid in the one good as a function of state  $s$  in period 2. Returns are not necessarily positive in all states.

Let us look at some special assets. First, the sure bond pays one unit of the good in each state. It is identified with  $j = 0$  and is defined by

$$(1a) \quad R_0(s) = 1 \quad \text{for all } s.$$

A second special case is trade in stocks. Let us identify home and foreign stocks (claims to home and foreign period 2 output, respectively) as assets  $j = h$  and  $j = f$ , defined by the return functions

$$(1b) \quad R_h(s) = y^2 \quad \text{and} \quad R_f(s) = y^{*2} \\ \text{for all } s.$$

Third, the Arrow-Debreu securities are the set of assets that each pay one unit of the good in one specific state only. We identify the Arrow-Debreu security for state  $s$  with  $j = s$ , for all  $s$ . It is defined by

$$(1c) \quad R_s(\sigma) = 1 \quad \text{for } \sigma = s, \\ R_s(\sigma) = 0 \quad \text{for all } \sigma \neq s$$

Let  $S$  be the (finite or infinite) number of different states of the world. In standard terminology, the asset market is said to be complete if the set  $J$  of assets is such that there are  $S$  linearly independent assets (that is, there are  $S$  linearly independent return

functions). Then agents can reach the same consumption bundle across states via trade in the available assets as they can via trade in the  $S$  Arrow-Debreu securities. If there are fewer than  $S$  linearly independent assets, the asset market is said to be incomplete. Our analysis does not presume that the asset market is complete or that trade in Arrow-Debreu securities is feasible, but incorporates these possibilities as special cases. Below, we shall sometimes assume that the state of the world is bivariate normally distributed. Then, whenever the number of assets is finite, the asset market is incomplete.

Let us now consider the home country. It has a representative consumer who is entitled to home output in the two periods. The consumer has a subjective probability distribution function  $F(s)$  over the states of the world. The consumer has preferences over period 1 consumption,  $c^1$ , and state-dependent period 2 consumption,  $c^2(s)$ . The preferences can be represented by the additively separable expected utility function

$$(2) \quad U(c^1) + \beta E[U(c^2)],$$

where  $U(\cdot)$  is a standard increasing concave sufficiently differentiable von Neumann-Morgenstern utility function,  $\beta > 0$  is the subjective discount factor, and  $E[x]$  denotes the subjective expected value  $\int x(s) dF(s)$ .<sup>9</sup>

Let  $m$  denote (net) import of period 1 goods, and let the  $J$  vector  $z = (z_j)_{j \in J}$  denote (net) import of the  $J$  assets from the world asset market in period 1. Then period 1 consumption and period 2 consumption in state  $s$  are given by<sup>10</sup>

$$(3a) \quad c^1 = y^1 + m \quad \text{and}$$

$$(3b) \quad c^2(s) = y^2 + \sum_{j \in J} R_j(s) z_j.$$

<sup>9</sup>As is well known, representing preferences by an additively separable expected utility function does not allow a separation between risk aversion and intertemporal substitution in consumption (see Larry Selden, 1978, 1979). When discussing differences in risk aversion, we shall actually use Selden's formulation to separate risk aversion from intertemporal substitution.

<sup>10</sup>We disregard bankruptcy issues, by not restricting consumption to be nonnegative.

It is practical to define preferences directly over import of period 1 goods and assets. Substitution of (3a, 3b) into (2) allows us to define the trade utility function  $\tilde{U}(m, z)$  by

$$(4) \quad \tilde{U}(m, z) = U(y^1 + m) + \beta E[U(y^2 + \sum_{j \in J} R_j(s) z_j)].$$

Let  $p$  and  $q = (q_j)_{j \in J}$  denote the price of period 1 goods and the  $J$  vector of asset prices. It is convenient to define the balance-of-payments (deficit) function  $B(p, q, u)$  as the minimum expenditure on import of goods and assets required to reach a given utility level. That is,

$$(5) \quad B(p, q, u) = \min \{ pm + qz \mid \tilde{U}(m, z) \geq u \},$$

where  $qz$  denotes the inner product  $\sum_{j \in J} q_j z_j$ . (The balance-of-payments function is simply the standard expenditure function minus the value of period 1 output.)<sup>11</sup>

In the rest of the paper we will take period 1 goods to be the numeraire,  $p = 1$ , and hence express asset prices  $q$  in terms of period 1 goods.

It is now easy to represent a *trade equilibrium* for the economy, an equilibrium in which the economy faces a given vector of asset prices  $q$  on the world asset market. It is simply given by the equations

$$(6) \quad B(1, q^t, u^t) = 0,$$

$$(7a) \quad m = B_p(1, q^t, u^t), \quad \text{and}$$

$$(7b) \quad z = B_q(1, q^t, u^t).$$

Equation (6) says that the balance of payments is zero in equilibrium, whereas equations (7a) and (7b) express import of goods and assets as the derivative of the balance-

of-payments function with respect to the price of period 1 goods and asset prices respectively, exploiting standard properties of expenditure functions. For given world asset prices  $q^t$ , equations (6) and (7) can be solved for the corresponding home utility level  $u^t$  and the import  $m$  and  $z$  of goods and assets.<sup>12</sup>

An *autarky equilibrium*, an equilibrium without access to the world asset market, is given by the equation (6) and

$$(8) \quad B_q(1, q, u) = 0,$$

the latter stating that the import of assets is zero. (Import of period 1 goods is then also zero,  $B_p(1, q, u) = 0$ , but by Walras's law that equation is redundant.) Equations (6) and (8) can be solved for the autarky asset prices  $q$  and the autarky utility level  $u$ .

It follows that the *gains-from-trade* theorem holds: Let  $u^t$  be the utility level associated with a trade equilibrium, and let  $u$  be the utility level in an autarky equilibrium. Then we have

$$(9) \quad u^t \geq u.$$

The proof is as in the standard trade model (see, for instance, Dixit and Norman, 1980, or Woodland, 1982). First, we have

$$(10) \quad B(1, q^t, u^t) = 0 \\ = m^a + q^t z^a \geq B(1, q^t, u).$$

The balance of payments in the trade equilibrium is zero (the first equality in (10)). This trivially equals the value at trade asset prices  $q^t$  of the autarky import  $m^a$  and  $z^a$  of period 1 goods and assets, since these are zero (the second equality in (10)). Zero im-

<sup>11</sup> This function occurs in the literature under a variety of names. See Peter Lloyd and Albert Schweinberger, 1986, for references to its use in previous literature.

<sup>12</sup> If the balance-of-payments function is not differentiable in  $p$  or  $q$ , goods and asset imports are not unique. We can then interpret  $B_p$  and  $B_q$  as correspondences. Our results below on the trade pattern do not depend on whether goods and asset imports are unique or not. For comparative statics of the Dixit and Alan Woodland (1982) type, it is necessary that the balance-of-payments function is differentiable.

port gives autarky utility level  $u$ . The minimum import expenditure at trade prices required to reach utility level  $u$  cannot be larger, and will be less if there is some substitution and trade prices differ from autarky prices (the inequality in (10)). Second, since the balance-of-payments function is increasing in utility, (9) follows from (10).

We note that the gains-from-trade theorem implies that trade in complete or incomplete asset markets is better than autarky. However, in analogy with the case with goods trade only, it does not follow that trade in more assets is better than in fewer, unless the prices of all previously traded assets remain unchanged. The usual terms-of-trade qualification applies: if the prices of assets previously imported (exported) increase (decrease) when trade in additional assets is opened up, the negative terms-of-trade effect may outweigh the gains from trade.

## II. World Equilibrium and the Law of Comparative Advantage

Next we shall consider a world equilibrium with trade between the home and foreign countries. The foreign country has access to a world market with the same set  $J$  of assets as the home country, a representative consumer entitled to foreign output in the two periods, and with a subjective probability distribution function  $F^*(s)$ , a von Neumann-Morgenstern utility function  $U^*(\cdot)$ , a subjective discount factor  $\beta^* > 0$ , and a trade utility function over period 1 goods (net) import  $m^*$  and asset (net) import  $z^*$ ,  $U^*(m^*, z^*)$ , defined by the analogue to (4). We can then represent the foreign country by a balance-of-payments function  $B^*(p, q, u^*)$  defined by the analogue of (5). A *trade equilibrium* for the foreign country is then, for given asset prices  $q^t$  relative to period 1 goods, the utility level  $u^{*t}$  and the import  $m^*$  and  $z^*$  of period 1 goods and assets that solve the equations' analogue to (6) and (7). An *autarky equilibrium* for the foreign country is an autarky asset price vector  $q^*$  and a utility level  $u^*$  that fulfill the analogues of (6) and (8).

A *world equilibrium* is a vector  $(q^t; m, z, u^t; m^*, z^*, u^{*t})$  such that  $(q^t, m, z, u^t)$  and

$(q^t, m^*, z^*, u^{*t})$  are trade equilibria for the home and the foreign country, respectively, and such that the world asset market and period 1 goods market are in equilibrium,

$$(11a) \quad z + z^* = 0 \quad \text{and}$$

$$(11b) \quad m + m^* = 0.$$

(The world market for period 1 goods is in equilibrium whenever the asset market is in equilibrium, given the budget constraint (6) for the home country and the analogue for the foreign country.)

Let  $m$  and  $z$  be the home country's import of period 1 goods and assets in a world equilibrium, and let  $q$  and  $q^*$  be home and foreign autarky asset prices relative to period 1 goods. Then the *law of comparative advantage* can be written on the form

$$(12) \quad (q - q^*)z \geq 0.$$

It states that on the average, the home country will import assets whose autarky prices are higher in the home country than in the foreign country. If only one asset is traded we have an exact relation between autarky asset prices and the trade pattern: The asset will be imported (and period 1 goods will be exported) if and only if the autarky price of the asset is higher in the home country than in the foreign country. If more than one asset is traded, the law of comparative advantage provides a "tendency" for a particular asset to be imported if its autarky price is relatively high, rather than an exact relation for import in any individual asset.<sup>13</sup>

<sup>13</sup>As Alan Deardorff (1980) emphasizes, a positive inner product  $xy = \sum_j x_j y_j \geq 0$  does not exactly provide a positive correlation between the  $J$  vectors  $x = (x_j)$  and  $y = (y_j)$ , unless either  $\sum_j x_j = 0$  or  $\sum_j y_j = 0$ . This is so, since the sample correlation coefficient  $\text{cor}(x, y)$  is proportional to the sample covariance  $\text{cov}(x, y)$  and the latter fulfills  $\text{cov}(x, y) = xy - \sum_j x_j \sum_j y_j / J$ . Deardorff shows how one can construct correlations in two ways. One way is to exploit the balance-of-payments constraint. Let  $q^t$  be the asset prices in terms of goods in the world equilibrium. Then (12) is equivalent to the statement that the  $(J+1)$  vectors  $(0, ((q_1 - q_j^*)/q_1^t))$  and  $(m, (q_j^t z_j))$  are positively correlated, since  $m + q^t z = 0$ . The other way is to restrict the vector of goods and

The proof of the law of comparative advantage is as in the standard trade model (see Deardorff, 1980; Dixit and Norman, 1980; or Woodland, 1982). We have

$$(13) \quad m + qz \geq B(1, q, u') \\ \geq B(1, q, u) = 0.$$

The first inequality follows since import  $(m, z)$  gives utility  $u'$  but is not necessarily the combination of net imports of goods and assets that minimize expenditure at autarky prices. The second inequality follows since we know from the gains-from-trade theorem that the home country's utility level  $u'$  in any trade equilibrium cannot fall short of the utility level in autarky  $u$ , and the balance-of-payments function is increasing in utility. The equality follows from the budget constraint (6). An analogous argument for the foreign country gives

$$(14) \quad m^* + q^*z^* \geq 0,$$

which we by (11a, b) can write as

$$(15) \quad -m - q^*z \geq 0.$$

Addition of (13) and (15) gives (12).

When discussing the determinants of the trade pattern, one can either examine the world equilibrium directly, or rely on the law of comparative advantage. In the former case, one discusses how differences between countries directly determine the trade pattern, without looking at the autarky prices. In the latter case, one looks at how differences between countries determine relative autarky

prices, and then from that indirectly infers the determinants of the trade pattern. In recent discussions of the trade pattern of goods and factors in the standard trade model, the former route has usually been chosen (see references mentioned in the introductory discussion). In our case, it is convenient to choose the latter route, since we can then directly apply a standard theory of asset pricing.

### III. The Pattern of Trade in Risky Assets

#### A. The Current Account and the Capital Account

Let us state the balance-of-payments relation for the home country in a trade equilibrium. We can write it as

$$(16) \quad m + q'z = B(1, q', u') = 0,$$

stating that the sum of the current account deficit (net import of goods  $m$ ) and the capital account deficit (the value of net import of assets  $qz$ ) is zero.<sup>14</sup> Hence what is being determined in a trade equilibrium is not only the aggregate current and capital account deficits, that is, whether the home country is a net borrower or lender (the intertemporal trade pattern), but also the components of the capital account, the disaggregate trade pattern in individual assets (the interstate trade pattern).

If we would like to concentrate on the intertemporal trade pattern, we could simplify the model by considering trade in only one asset, and even disregard the effect of uncertainty and incomplete markets by then assuming that there is no uncertainty and only one state in period 2. This gives us the simplest possible model to discuss international borrowing and lending. If we would like to concentrate exclusively on the trade pattern in risky assets, we could eliminate the first period, and assume that assets are traded before uncertainty is resolved. This

asset prices to be in the unit simplex. Let  $(p, q)$  and  $(p^*, q^*)$  be the home and foreign autarky prices of period 1 goods and assets. The proof in the next paragraph of the text gives  $(p - p^*, q - q^*)(m, z) \geq 0$ . Restricting  $(p, q)$  and  $(p^*, q^*)$  to be in the unit simplex then implies that the  $(J+1)$  vectors  $((1, q)/(1 + \sum_j q_j) - (1, q^*)/(1 + \sum_j q_j^*))$  and  $(m, z)$  are positively correlated.

For our purpose it is sufficient to interpret (12) as stating that there is tendency for asset  $j$  to be imported into the home country ( $z_j > 0$ ) when its home autarky price (measured in goods) is higher than its foreign autarky price (measured in goods) ( $q_j > q_j^*$ ).

<sup>14</sup>Since there is no initial international debt, the trade balance and the current account coincide.

then abstracts from intertemporal trade and gives us the simplest possible model of trade in risky assets, "interstate" trade.

As we shall see, in the more general model intertemporal trade and interstate trade are not independent, and, for instance, the available assets affect a country's current account. Therefore, we choose to keep the two-period framework. This also has the advantage that the expressions for asset prices to be derived are similar to those used in the asset-pricing literature.

### B. Autarky Asset Prices

The home autarky asset price  $q_j$  of a particular asset  $j$  with return function  $R_j(s)$  is simply given by the marginal rate for substitution between asset  $j$  and period 1 goods of the trade utility function (4) at zero import of goods and assets,  $\tilde{U}_j(0,0)/\tilde{U}_m(0,0)$ , where  $\tilde{U}_j$  and  $\tilde{U}_m$  denote the partial with respect to  $z_j$  and  $m$ . It follows from (4) that the autarky asset price will fulfill

$$(17) \quad q_j = \beta E[U_c(y^2)R_j]/U_c(y^1),$$

the familiar expression of the discounted expected utility of period 2 returns over the marginal utility of period 1 consumption.

It is practical to relate the price of an asset to the real interest rate on a sure bond, and to the risk measure for the asset. First, define the autarky real interest rate,  $r$ , from the autarky asset price on the sure bond,

$$(18) \quad q_0 = 1/(1+r) \\ = \beta E[U_c(y^2)]/U_c(y^1),$$

where we have substituted (1a) in (17). Second, let us define the autarky risk measure for asset  $j$ ,  $\Pi_j$ , as

$$(19) \quad \Pi_j = -\text{Cov}[U_c(y^2), R_j]/E[U_c(y^2)].$$

Third, use the rule  $E[xy] = E[x]E[y] + \text{Cov}[x, y]$  to rewrite (17), and apply the definitions (18) and (19). This gives,

$$(20) \quad q_j = \{E[R_j] - \Pi_j\}/(1+r).$$

We see that the asset price can be written as the present value of the difference between its expected return and its risk measure.

The risk measure is proportional to the negative of the covariance between the marginal utility of period 2 consumption  $U_c(y^2(s))$  and the returns  $R_j(s)$ .<sup>15</sup> Hence it is positive or negative depending upon whether period 2 marginal utilities and returns are negatively or positively correlated. The risk measure for an asset can be interpreted as a measure of how risky that asset is relative to the sure bond. If the risk measure is positive, the asset is riskier than the sure bond. If it is negative, the asset is less risky than the sure bond.<sup>16</sup>

It is clear from (20) that autarky prices for a given asset may differ across countries because autarky interest rates, autarky risk measures, or both, differ across countries. If the subjective beliefs, the subjective probability distributions over states of the world, differ across countries, autarky asset prices may differ also because the expected return for a given asset differs. The analysis below consequently examines the underlying determinants of differences in autarky interest rates, risk measures, and expected returns.

### C. Trade in Risky Assets

We shall examine the difference between the home and foreign countries' autarky asset prices of a given asset  $j \in J$ . We will look for conditions under which the home country's autarky asset price exceeds the foreign country's autarky asset price, and

<sup>15</sup>The risk premium can be defined as the difference between the expected gross rate of return,  $E[R_j]/q_j$ , and the gross real rate of interest,  $1+r$ . Then the risk premium is equal to  $\Pi_j/q_j$  and fulfills  $\Pi_j/q_j = -\beta \text{Cov}[U_c(y^2)/U_c(y^1), R_j/q_j]$  and is hence the negative of the covariance between the marginal rates of substitution and the *ex post* rates of return  $R_j(s)/q_j$ .

<sup>16</sup>Note that the sure bond has a sure return, but that the utility value of the return is risky, since marginal utility itself is risky. Hence there is nothing paradoxical with assets that are less risky than the sure bond. A sure-utility bond (in autarky) ( $j=u$ ) would have returns  $R_u(s)$  fulfilling  $U_c(y^2)R_u(s)=1$ , hence  $R_u(s)=1/U_c(y^2)$  for all  $s$ .



hence under which there will be a tendency in a world equilibrium for asset  $j$  to be imported by the home country and exported by the foreign country. In the special case where asset  $j$  is the only traded asset we will know for sure that asset  $j$  will be imported.

The home autarky asset price of asset  $j$  is given by expression (17) or (20). The foreign autarky asset price is given by an analogous expression, with an asterisk denoting foreign output and preferences. Let us now assume that the subjective probability distribution is the same in the home and foreign country,

$$(21) \quad F(s) = F^*(s) \quad \text{for all } s,$$

so the expected return for a given asset  $j$  is the same in both countries,

$$(22) \quad E[R_j] = E^*[R_j].$$

(Below we shall discuss also the case when the subjective probability distribution differs across countries and (22) does not hold.) Let us also restrict the discussion to assets with positive expected return,

$$(23) \quad E[R_j] > 0.$$

(If the expected return is negative, we can simply redefine the asset by changing the sign of its returns.)

If the countries are identical in all respects, the autarky asset prices will be identical, there is no basis for trade, and zero trade will be a trade equilibrium. Hence, trade here arises because of differences between the countries. The countries can differ either with regard to their outputs, or with regard to their preferences, including their subjective probability distributions. Let us first consider a situation when the only difference between the countries is with regard to their outputs.

*Differences in Output.* Thus, we assume that the foreign country is identical to the home country in all respects except the outputs, and we drop the asterisk on the foreign country's preferences.

Let us first look for conditions under which the home autarky interest rate is lower than

the foreign one,

$$(24) \quad r < r^*.$$

A lower home autarky interest rate implies by (20) that for all assets, which do not have higher autarky risk measures at home than abroad, home autarky prices will be higher, and there is a tendency for the home country to import all such assets. For assets with a higher autarky risk measure at home, a lower home autarky real interest rate implies a higher autarky price but not necessarily higher than the foreign autarky price. Nevertheless, we may state that a lower home autarky rate contributes to a tendency to import all assets into the home country, to run a home capital account deficit, and hence for the home country to be a net lender. This is true also if the sure bond does not exist. If the *only* asset traded is the sure bond, we have an exact result and know for sure that the home country will import the sure bond and be a net lender.

We can examine this by looking at the difference in autarky prices of the sure bond. The difference is given by

$$(25) \quad q_0 - q_0^* = [1/(1+r) - 1/(1+r^*)] \\ = \beta E \left[ (U_c(y^2)/U_c(y^1)) \right. \\ \left. - (U_c(y^{*2})/U_c(y^{*1})) \right].$$

We would like to know under what conditions this difference is positive. Let us first assume that the countries differ only with respect to period 1 output. We then have

$$(26) \quad q_0 - q_0^* = \beta E [U_c(y^2)] \\ / [1/U_c(y^1) - 1/U_c(y^{*1})].$$

Since the marginal utility of consumption is decreasing, it follows directly that the home autarky price of the sure bond is higher, and the home autarky interest rate lower, if the home country has a higher period 1 output,

$$(27) \quad y^1 > y^{*1}.$$

This is a standard consumption smoothing result (across countries, though, not across time).<sup>17</sup> The home country has relatively more output in period 1, and it will export goods in period 1 and import goods in period 2, by being a net lender in period 1.

Let us next assume that period 1 output is the same in the two countries, but that period 2 output is different. Then we have

$$(28) \quad q_0 - q_0^* = \beta E[U_c(y^2) - U_c(y^{*2})] / U_c(y^1).$$

Since the marginal utility of consumption is decreasing, it follows (see Theorem 1 in Steven Lippman and John McCall, 1981) that a sufficient condition for (28) to be positive is that home period 2 output is stochastically smaller than foreign period 2 output, that is, home period 2 output is first-order stochastically dominated by foreign period 2 output, denoted

$$(29) \quad y^2 <_1 y^{*2}.$$

First-order stochastic dominance of home output by foreign output implies that the expected value of home output is smaller,

$$(30) \quad Ey^2 > Ey^{*2},$$

and can be understood as a generalization of that property.<sup>18</sup>

This result can also be interpreted as a straightforward consumption smoothing result. If the home country has lower expected period 2 output than the foreign country, it will export goods in period 1 and import

goods in period 2, by being a net lender in period 1.

Under the assumption that preferences exhibit nonincreasing absolute risk aversion the third-order derivative  $U_{ccc}$  of the von Neumann-Morgenstern utility function is positive,<sup>19</sup>

$$(31) \quad U_{ccc} > 0,$$

and the marginal utility of consumption is a convex function of consumption. Then, another sufficient condition for (28) to be positive (see Theorem 2 in Lippman and McCall, 1981) is that home period 2 output is more risky than foreign period 2 output, that is, home period 2 output is second-order stochastically dominated by foreign period 2 output, denoted

$$(32) \quad y^2 <_2 y^{*2}.$$

A special case of this is when home and foreign period 2 output have the same mean but home output has a larger variance,

$$(33) \quad \text{Var}[y^2] > \text{Var}[y^{*2}],$$

or when home period 2 output is a mean-preserving spread of foreign period 2 output. Second-order stochastic dominance can be understood as a generalization of those special cases.<sup>20</sup>

Intuitively we can understand this result the following way. If marginal utility is a convex function of consumption, Jensen's inequality implies that increased variance in consumption increases expected marginal utility, which increases the price of the sure bond and decreases the interest rate. If the

<sup>17</sup>If both countries have less period 1 output than period 2 output (average or for each state of the world), home consumption becomes more unevenly divided over time with trade in the sure bond than in autarky.

<sup>18</sup>Let  $G(\cdot)$  and  $G^*(\cdot)$  denote the cumulative distribution functions for the random variables  $y$  and  $y^*$ , respectively. We say that  $y^*$  is stochastically larger than  $y$ , written  $y^* >_1 y$ , or  $G^* >_1 G$ , if and only if  $G(x) - G^*(x) \leq 0$  for all  $x$ . Equivalently, we say that  $y^*$  stochastically dominates  $y$  to the first order. See Lippman and McCall (1981).

<sup>19</sup>The measure of local absolute risk aversion is  $-U_{cc}/U_c$ . We have  $(d/dc)(-U_{cc}(c)/U_c(c)) = -U_{ccc}/U_c + (U_{cc}/U_c)^2 \leq 0$ , which implies  $U_{ccc} \geq (U_{cc})^2/U_c > 0$ .

<sup>20</sup>Let  $G(\cdot)$  and  $G^*(\cdot)$  denote the cumulative distribution functions for the random variables  $y$  and  $y^*$ , respectively. We say that  $y^*$  is less risky than  $y$ , written  $y^* >_2 y$ , or  $G^* >_2 G$ , if and only if  $\int_{-\infty}^x [G(z) - G^*(z)] dz \geq 0$  for all  $x$ . Equivalently, we say that  $y^*$  stochastically dominates  $y$  to the second order. See Lippman and McCall (1981).

TABLE 1.—SUMMARY OF RESULTS

(Import (Yes or No (= Export), or Condition for Import) of What Asset Under What Difference)				
Asset:	Sure bond ( $r < r^*$ )	Arbitrary asset ( $\Pi_j < \Pi_j^*$ )	Stocks	Arrow-Debreu
Differences in:				
(i) Output	$y^1 > y^{*1}$ $y^2 <_1 y^{*2}$ $y^2 <_2 y^{*2}$	$\text{Cov}[U_c(y^2), R_j] > \text{Cov}[U_c(y^{*2}), R_j]$ $\text{Cov}[y^2, R_j] < \text{Cov}[y^{*2}, R_j]$	$h$ : No $f$ : Yes	$y^2 < y^{*2}$
(ii) Time Preference ( $\beta > \beta^*$ )	Yes	Yes	Yes	Yes
(iii) Risk Aversion ( $\gamma > \gamma^*$ )	Yes	$\text{Cov}[y^2, R_j] < 0$ $\Pi_j < 0$	No	$y^2 < \bar{y}^2$
(iv) Subjective Beliefs ( $F \neq F^*$ )	$F <_1 F^*$ $F <_2 F^*$	$\int (f(s) - f^*(s)) U_c(s) R_j(s) ds > 0$	$\rho > 1$ : $F <_1 F^*$ or $F <_2 F^*$ $\rho < 1$ : $F >_1 F^*$ or $F >_2 F^*$	$f(s) > f^*(s)$

third-order derivative is negative, the opposite result holds. This is an example of the ambiguity of the effect on saving on increased riskiness of future income (see the survey by Agnar Sandmo, 1974). In the literature there is a general agreement that non-increasing absolute risk aversion and hence a positive third-order derivative of the von Neumann-Morgenstern utility function is the most relevant case. Thus, the country with the riskier period 2 output will have a tendency to import the sure bond, and having the riskier period 2 outputs contributes to a tendency to import all assets and be a net lender.

The results above for the sure bond are summarized in Table 1, row (i), first column.

Let us next turn to differences in the risk measures. From (20) we see that, for a home autarky real interest not higher than the foreign one, a lower risk measure at home for asset  $j$  implies a higher home autarky asset price and hence a tendency for asset  $j$  to be imported into the home country. For a home autarky interest rate higher than the foreign one, a lower home autarky risk measure implies a higher autarky asset price, but not necessarily higher than in the foreign country. Risk measures are specific to individual assets and depend on the individual risk characteristics of the assets. Hence a

difference in risk measures for a given asset gives information about trade in that specific asset; a difference in autarky real interest rates affect autarky asset prices for all assets and hence gives information about aggregate asset trade, the capital account.

Let us assume that autarky interest rates are the same, in order to focus on differences in autarky risk measures alone. Let us look at conditions for the home autarky risk measure for asset  $j$  to be lower than the foreign one,

$$(34) \quad \Pi_j < \Pi_j^*.$$

We assume that period 1 output is the same in both countries. From (18) and equal autarky interest rates it follows that  $E[U_c(y^2)] = E[U_c(y^{*2})]$ . Then, from (19) we see that the home autarky risk measure then is lower, if and only if

$$(35) \quad \text{Cov}[U_c(y^2), R_j] > \text{Cov}[U_c(y^{*2}), R_j],$$

that is, if the return is more positively correlated with home marginal utility of consumption than with foreign marginal utility of consumption.

Since marginal utility of consumption is decreasing in consumption, we might believe that (35) is equivalent to the simple condition that the return should be more negatively correlated with home period 2 output than with foreign period 2 output,

$$(36) \quad \text{Cov}[y^2, R_j] < \text{Cov}[y^{*2}, R_j].$$

This is so only in special cases, though. One interesting special case is when the von Neumann-Morgenstern utility function has constant absolute aversion, that is, when

$$(37) \quad U(c) = -e^{-\gamma c},$$

with the constant  $\gamma = -U_{cc}/U_c > 0$  being Arrow-Pratt's measure of absolute risk aversion. If in addition period 2 outputs and asset return are all jointly normally distributed,<sup>21</sup> it is easy to apply a theorem by Mark Rubinstein (1976)<sup>22</sup> and show that the risk measure is simply given by

$$(38) \quad \Pi_j = \gamma \text{Cov}[y^2, R_j].$$

Then, for  $\gamma = \gamma^*$ ,  $\Pi_j < \Pi_j^*$  is equivalent to (36).<sup>23</sup>

We conclude that, under the assumption of equal autarky interest rates, the condition is simply that the return should be more negatively correlated with home period 2 output than with foreign period 2 output. Then asset  $j$  is less risky in the home country, its autarky risk measure is lower, its autarky asset price is higher, and there is a tendency that the asset will be imported by the home country. This result is reported in Table 1, row (i), the second column.

<sup>21</sup>Note that, as usual, the assumption of a normal distribution of outputs is problematic, since it implies that outputs can take negative values with positive probability.

<sup>22</sup>The theorem says that, if  $x$  and  $y$  are bivariate normal, under some mild regularity conditions,

$$\text{Cov}[f(x), y] = E[f_x(x)] \text{Cov}[x, y].$$

<sup>23</sup>Other cases when (35) and (36) are approximately equivalent are discussed in a previous working paper version of this paper.

Let us next consider trade in home and foreign stocks. Because of symmetry we need only look at foreign stocks. Trade in foreign stocks is of course affected by differences in autarky interest rates, since these affect trade in all assets. Suppose now that autarky interest rates are the same. Then the autarky risk measure is the only source of differences in autarky asset prices. The condition for the home autarky risk measure for the foreign stocks to be low, and thus for a tendency for the home country to import foreign stocks, is, from (36),

$$(39) \quad \text{Cov}[y^2, y^{*2}] < \text{Cov}[y^{*2}, y^{*2}] \\ = \text{Var}[y^{*2}].$$

We know that

$$\text{Cov}[y^2, y^{*2}] \leq (\text{Var}[y^2] \text{Var}[y^{*2}])^{1/2}.$$

If we assume that home and foreign period 2 output has the same variance, which from our previous discussion is in accordance with the assumption of equal autarky interest rates, we get that a sufficient condition for (34) is that home and foreign outputs are less than perfectly positively correlated. Thus, there is a tendency for the home country to import foreign stocks if the two period 2 outputs are not perfectly correlated. By symmetry, there will be a tendency for the home country to export home stocks, if home and foreign period 2 outputs are less than perfectly correlated. These results are summarized in Table 1, row (i), third column.

Let us finally consider Arrow-Debreu securities. Let  $f(s)$  denote either the probability of state  $s$  (if the probability distribution is discrete) or the probability density for state  $s$  (if the probability distribution is absolute-continuous). From (1c) and (17) the home autarky price of Arrow-Debreu security  $s$ , for all  $s$ , will then be

$$(40) \quad q_s = \beta f(s) U_c(y^2) / U_c(y^1).$$

When the countries' period 1 outputs are equal, it follows directly that there is a tend-

ency for Arrow-Debreu security  $s$  to be imported if home-period 2 output is state  $s$  is lower than that of the foreign country,

$$(41) \quad y^2 < y^{*2}.$$

That is, trade in Arrow-Debreu securities is simply related to the relative scarcity of period 2 output.

*Summary.* Table 1, row (i), summarizes the results on output differences and asset trade. First, in general a low home autarky interest rate contributes to a tendency for the home country to import all assets and be a net lender. If the only traded asset is a sure bond, it will definitely be imported by the home country. The home autarky interest rate is low if home period 1 output is high, or if home period 2 output is stochastically smaller than foreign period 2 output. The home autarky interest is also low if preferences exhibit nonincreasing absolute risk aversion, and if home period 2 output is riskier than foreign period 2 output. Second, in general a low autarky risk measure for an asset (the product of the risk premium and the asset price) contributes to a tendency for the home country to import the asset. The autarky risk measure is low if the asset's returns are more positively correlated with home autarky period 2 marginal utility than with foreign autarky period 2 marginal utility. If autarky interest rates are equal, under some restrictions there is a more specific result: If the joint probability distributions between returns and period 2 outputs are normal and there is constant absolute risk aversion, the autarky risk measure is low if the asset's return is more negatively correlated with home output than with foreign output. Third, if autarky interest rates are equal, there is a tendency for the home country to import foreign stocks, and export home stocks, if home and foreign outputs are less than perfectly positively correlated. Fourth, there is a tendency to import an Arrow-Debreu security for a particular state if home period 2 output in that state is lower than in the foreign country.

Next, we assume that outputs are identical in the two countries, but that preferences

differ.<sup>24</sup> We shall consider differences in the rate of time preference (the subjective discount factor), the degree of risk aversion, and the subjective probability distribution.

*Differences in the Rate of Time Preference.* The effect of differences in the rate of time preference is easy to see. Consider the situation when the home country has a lower rate of time preference than the foreign country. That is, the home subjective discount factor is larger,

$$(42) \quad \beta > \beta^*.$$

It follows directly from the definition of the autarky asset price (17) that home autarky asset prices will be higher for all assets (with positive asset prices).<sup>25</sup>

*Summary.* When the home country has a lower rate of time preference, there is a tendency for all assets to be imported into the home country, and for the home country to be a net lender.

*Differences in Risk Aversion.* We would like to consider differences in risk aversion across countries. This is a bit problematic with expected utility preferences like (2), since in that formulation attitudes toward risk cannot be separated from intertemporal substitution. Therefore, we choose to use a formulation according to Selden (1978), which allows such a separation. More precisely, we assume that there are *intertemporal preferences* over period 1 consumption,  $c^1$ , and *certainty equivalent* period 2 consumption,  $\hat{c}^2$ , according to the intertemporal utility function

$$(43) \quad U(c^1) + \beta U(\hat{c}^2).$$

<sup>24</sup>We assume  $y^2 = y^{*2}$ , that is, home and foreign period 2 output are identical and hence perfectly correlated. This is of course not equivalent to assuming that home and foreign output are i.i.d. In the former case, claims to home and foreign output are perfect substitutes. In the latter case, they are not.

<sup>25</sup>We realize from (20) that assuming that expected dividends are positive, (23), is not the same thing as assuming that the asset price is positive, since the risk term may be positive and larger than the present value of the expected return.

Attitudes toward risk are represented by the risk utility function  $V(c^2)$ , by which the certainty equivalent period 2 consumption is defined according to

$$(44) \quad V(\hat{c}^2) = E[V(c^2(s))], \quad \text{or} \\ \hat{c}^2 = V^{-1}\{E[V(c^2(s))]\}.$$

We restrict preferences to have constant absolute risk aversion, that is,

$$(45) \quad V(c^2) = -e^{-\gamma c^2}.$$

The trade utility function is defined by  $\tilde{U}(m, z) = U(y^1 + m) + \beta U\{V^{-1}\{E[V(y^2 + \sum_{j \in J} R_j(s)z_j])]\}$ . It can then be shown that autarky asset prices  $q_j = \tilde{U}_j(0, 0)/\tilde{U}_m(0, 0)$  can still be written as in (20), with the risk measure defined by (19). With the risk utility function fulfilling (45), under the assumption that period 2 output and asset returns are jointly normally distributed, the risk measure is indeed given by (38). The difference is that the autarky price of sure bonds and the autarky real interest rate are given by

$$(46) \quad q_0 = 1/(1+r) = \beta U_c(\hat{y}^2)/U_c(y^1),$$

where the certainty equivalent period 2 output is given by<sup>27</sup>

$$(47) \quad \hat{y}^2 = E[y^2] - \gamma \text{Var}[y^2]/2.$$

We now assume that the countries differ only with respect to the measure of absolute risk aversion, and that the home country is more risk averse,

$$(48) \quad \gamma > \gamma^*.$$

First, we examine interest rates. We see immediately from (47) that when the home country is more risk averse, the home cer-

tainty equivalent period 2 output,  $\hat{y}^2$ , will be lower than the foreign one,  $\hat{y}^{*2} = E[y^2] - \gamma^* \text{Var}[y^2]/2$  (although home and foreign period 2 outputs are identical). Therefore, the home autarky price of the sure bond will be higher and the real interest rate will be lower,

$$(49) \quad r > r^*.$$

This contributes to a tendency for the more risk-averse home country to import all assets.

Next, we look at the autarky risk measures for a given asset  $j$ . In order to ensure that differences in autarky risk measures are the only reason for trade, we assume that autarky real interest rates are equal. Since, as we have seen above, autarky real interest rates differ between the countries, when the home country is more risk averse and their intertemporal preferences are identical, we now assume that the subjective discount factors differ so as to equalize the autarky real interest rates.

The difference in the autarky risk measures equals, by (38),

$$(50) \quad \Pi_j - \Pi_j^* = (\gamma - \gamma^*) \text{Cov}[y^2, R_j].$$

It follows from (48) that the condition for the home autarky risk measure to be lower, and for a tendency for asset  $j$  to be imported into the home country, is

$$(51) \quad \text{Cov}[y^2, R_j] < 0.$$

The return should be negatively correlated with period 2 output. From (38) and (51) this also implies that the risk measure should be negative,

$$(52) \quad \Pi_j < 0.$$

Since the sure bond has a zero risk measure, this means that the asset should be less risky than the sure bond. Thus there is a tendency for the more risk-averse home country to import assets which are less risky than the sure bond.

<sup>26</sup>Note that when the intertemporal utility function is identical to the risk utility function,  $U(\cdot) = V(\cdot)$ , (37) and (43) imply the expected utility preference (2).

<sup>27</sup>We use that for  $x$  normally distributed,  $E[e^{-\alpha x}] = \exp[-\alpha(E[x] - \alpha \text{Var}[x]/2)]$ .

Consider also trade in stocks (claims to period 2 output,  $R_h(s) = R_f(s) = y^2$ ). Since period 2 output is positively correlated with itself, it follows directly from the above analysis that there is a tendency for stocks to be exported by the more risk-averse home country, since they have a positive risk measure and are riskier than the sure bond.

Let us finally consider the special case of Arrow-Debreu securities. From (1c) and the definition of the trade utility function it follows that the home autarky price of a particular Arrow-Debreu security  $s = y^2$  (since home and foreign period 2 outputs are now identical, the state can simply be identified with the period 2 output in each country) is given by

$$(53) \quad q_s = [f(y^2)V_c(y^2)/V_c(\hat{y}^2)]/(1+r).$$

Assuming that autarky interest rates are equal and using (45) and (47) and some algebra gives the ratio between home and foreign autarky prices of the security,

$$(54) \quad q_s/q_s^* = \exp\left\{-(\gamma - \gamma^*)\left[y^2 - \left(E(y^2) - (\gamma + \gamma^*)\text{Var}(y^2)/2\right)\right]\right\}.$$

It follows that there is a tendency for the security to be imported for states for which period 2 output falls short of a given level of period 2 output  $\bar{y}^2 = E(y^2) - (\gamma + \gamma^*)\text{Var}(y^2)/2$ . When period 2 output is sufficiently low, marginal utility in the home country is higher since a higher risk aversion means that marginal utility decreases more rapidly with consumption.

*Summary:* The results under the assumption that the home country has a higher constant absolute risk aversion than the foreign country are summarized in Table 1, row (iii). First, when home and foreign intertemporal preferences are identical, the home autarky interest rate is lower, which contributes to a tendency for the home country to import all assets and be a net lender. Second, when also subjective discount factors differ so as to make autarky real interest rates equal, there is a tendency for the more risk-averse home country to import assets

with negative risk measures, that is, assets that are negatively correlated with period 2 output and less risky than the sure bond. Third, there is then a tendency for the home country to export stocks, since they are assets which are more risky than the sure bond. Fourth, there is a tendency for the home country to import Arrow-Debreu securities for states with sufficiently low period 2 output.

*Differences in Subjective Beliefs.* Finally, we consider the case when countries differ only with respect to their subjective probability distributions, their beliefs. That is, their subjective probability distributions are no longer identical,

$$(55) \quad F(s) \neq F^*(s).$$

For a given asset  $j$  with returns  $R_j(s)$  it is no longer true that  $E[R_j(s)] = \int R_j(s) dF(s)$  is equal to  $E^*[R_j(s)] = \int R_j(s) dF^*(s)$ . Therefore, the previous method of expressing the asset price in terms of the real interest rate and the risk measure is not applicable. It is no longer true that a low autarky interest rate increases the relative autarky price for all assets. Hence it is no longer true that a low autarky interest rate contributes to a tendency for all assets to be imported. A low autarky interest rate implies only that there is a tendency for the sure bond to be imported.

Assume that preferences are again represented by the expected utility function (2).<sup>28</sup> From (17) it follows that the difference between the autarky prices of the sure bond is

$$(56) \quad q_0 - q_0^* = \beta \left\{ E[U_c(y^2)] - E^*[U_c(y^2)] \right\} / U_c(y^1).$$

We can directly apply our results on the autarky interest rates for differences in period 2 output. First, since marginal utility of consumption is decreasing, as sufficient condi-

<sup>28</sup> That is, the risk utility function in (43) is assumed to be identical to the intertemporal utility function in (44).

tion for a lower home autarky interest rate is that the home subjective probability distribution over (both countries') period 2 output,  $F(y^2)$ , (recall that  $s = y^2$ ) is first-order dominated by the foreign subjective probability distribution over (both countries') period 2 output,  $F^*(y^2)$ , that is,

$$(57) \quad F <_1 F^*.$$

Put differently, the home country has more pessimistic beliefs about both countries' period 2 output than the foreign country. Second, if the von Neumann-Morgenstern utility function has nonincreasing absolute risk aversion, marginal utility is convex, and a sufficient condition for a lower home autarky interest rate is that the home subjective probability distribution over (both countries') period 2 output is second-order dominated by the foreign subjective probability distribution over (both countries') period 2 output, that is,

$$(58) \quad F <_2 F^*.$$

Put differently, the home country believes that both countries' period 2 output is more risky than the foreign country believes.

For an arbitrary asset  $j$ , the difference between the home and foreign autarky price of asset  $j$  is

$$(59) \quad q_j - q_j^* = \beta \int (f(y^2) - f^*(y^2)) \times U_c(y^2) R_j(y^2) dy^2 / U_c(y^1)$$

(when the distributions are absolute-continuous; the analogue for discrete distributions is obvious). Expression (59) states that there is a tendency for asset  $j$  to be imported into the home country if the probability density differences,  $f(s) - f^*(s)$ , are positively correlated with the marginal-utility weighted returns,  $(U_c(y^2) R_j(s))$ .<sup>29</sup> Thus, we have the

rather obvious result that the home country has a tendency to import an asset when it assigns higher probabilities than the foreign country to the states where the assets pay well (where paying well means that the product of marginal utility of consumption and returns is large).

For stocks, the autarky price difference is

$$(60) \quad q_h - q_h^* = \beta \{ E[U_c(y^2)y^2] - E^*[U_c(y^2)y^2] \} / U_c(y^1).$$

Let us consider the case with constant relative risk aversion  $\rho$  (and intertemporal elasticity of substitution  $1/\rho$ ),<sup>30</sup>

$$(61) \quad U(c) = c^{1-\rho} / (1-\rho), \quad \rho > 0.$$

We have that the product of marginal utility and output is  $U_c(y^2)y^2 = (y^2)^{1-\rho}$ . This product is increasing or decreasing depending upon whether the degree of relative risk aversion is below or above unity.

Let us consider the case when the degree of relative risk aversion is above unity ( $\rho > 1$ ). Then the product of marginal utility and output is decreasing and convex, and we have the same two sufficient conditions for a tendency for the home country to import stocks as we have stated above for the tendency to import the sure bond, namely that the home country has more pessimistic beliefs about both countries' period 2 output than the foreign country (57), or that home country believes that both countries' period 2 output is more risky than the foreign country believes (58).

If the degree of relative risk aversion is below unity ( $\rho < 1$ ), the product of marginal utility and output is increasing and concave. Then the two sufficient conditions are reversed. The home country should have more optimistic beliefs about both countries' period 2 output than the foreign country, that is,

$$(62) \quad F >_1 F^*,$$

<sup>29</sup>We note that (59) being positive is equivalent to a positive correlation between the  $f(s) - f^*(s)$  and  $U_c(y^2) R_j(s)$ , since  $\int (f(s) - f^*(s)) ds = 0$  (compare fn. 13 above).

<sup>30</sup>In terms of Selden's formulation,  $V(\cdot)$  and  $U(\cdot)$  are identical and given by (61).



or the home country should believe that both countries' period 2 output is less risky than the foreign country believes, that is,

$$(63) \quad F >_2 F^*.$$

For the special case of Arrow-Debreu securities, the difference in autarky prices for security  $s = y^2$  is simply

$$(64) \quad q_s - q_s^* = \beta(f(y^2) - f^*(y^2)) \\ \times U_c(y^2)/U_c(y^1).$$

We see that there is a tendency to import Arrow-Debreu securities for states that are assigned larger probabilities by the home country

$$(65) \quad f(y^2) > f^*(y^2).$$

*Summary.* The results on differences in subjective beliefs are summarized in Table 1, row (iv). First, the home autarky interest rate will be low, and there will hence be a tendency for the home country to import the sure bond, if the home country has more pessimistic beliefs about the two countries' period 2 output than the foreign country, or (when preferences in the two countries exhibit nonincreasing absolute risk aversion) the home country believes that both countries' period 2 output is more risky than the foreign country believes. Counter to previous cases, a low home autarky interest rate does not imply that home autarky prices for other assets are low, and hence does not necessarily contribute to a tendency to import all assets. Second, there is, rather obviously, a tendency for the home country to import an arbitrary asset if the home country assigns higher probabilities than the foreign country to states for which the marginal utility times returns is high. Third, the tendency to import stocks (claims to period 2 output) depends on the degree of *relative* risk aversion. If the degree of relative risk aversion is above (below) unity, there is a tendency for the home country to import a claim to period 2 output if the home country has more pessimistic (optimistic) beliefs about the two

countries' period 2 output than the foreign country, or if the home country believes the two countries' period 2 output is more (less) risky than the foreign country. Fourth, there is a tendency to import Arrow-Debreu securities for states (period 2 output levels) that are assigned higher probabilities by the home country than by the foreign country.

#### IV. Conclusions

We have presented a theory of the determinants of the trade pattern in risky assets, by extending the law of comparative advantage according to which trade is correlated with autarky price differences. Hence we have looked at how differences between countries with regard to technology, endowments, and preferences determine autarky asset price differences and consequently the trade pattern in risky assets. We have derived results on the effect of differences in (i) output/endowments, (ii) rate of time preferences, (iii) risk aversion, and (iv) subjective beliefs on the trade pattern in arbitrary risky assets as well as the special cases of sure bonds, stocks, and Arrow-Debreu securities. The results have been summarized in highlighted paragraphs at the end of each subsection of Section III, and they are also summarized in Table 1.

We realize from our results that, when asset markets are incomplete, overall capital account deficits or surpluses depend on what assets are available for international trade. For instance, consider the case when countries differ only with respect to the stochastic properties of their output. If there is trade in claims to one country's output only, whether a country is a net borrower or lender depends on whether it has claims to its output or another country's output that is traded (as we saw above, a country has a tendency to export claims to its own output and import claims to other countries' output). It follows that in a monetary model with incomplete markets, it will matter for the capital flows what currency available assets are nominated in, since the real return on the assets will be affected by price-level risk.

The results derived have been interpreted in terms of trade in risky assets between

countries. Obviously, the model and its results can also be interpreted in terms of trade in risky assets between individuals.<sup>31</sup>

An important simplifying characteristic of our approach is that an asset is defined in terms of an exogenously given vector of next period's *gross real* returns across states of the world. We share this characteristic with most of the finance literature. Most assets, however, have gross real returns endogenously determined. For instance, the returns on equity, being claims to profits, are clearly endogenously determined when production decisions and goods and factor prices are endogenously determined. Even for an asset with exogenously given returns in terms of a particular good, the appropriate "real" return depends on endogenous relative goods prices when there are many goods. With many periods, the gross return in next period on a long-term asset is the sum of next period's endogenous asset price and the "direct" return/dividend (which may or may not also be endogenous). Generally, for most assets the stochastic properties of the gross real returns are endogenously determined and part of the equilibrium, and the stochastic properties differ between trade equilibria and autarky equilibria. From the point of view of our approach, if an asset has one gross real return vector in a trade equilibrium, and another gross real return vector in autarky, it is actually two different assets.

Hence, since most assets have endogenous gross returns, it may seem that our approach with exogenously specified gross returns should have very restricted applicability. We argue, however, that our approach can be used also to predict the trade pattern for assets with endogenously determined returns. The trick is to identify a particular asset's (endogenously determined) gross real

return vector across states of the world *in a trade equilibrium*, and then ask how a hypothetical asset with such a gross real return vector (taken to be exogenous and hence held fixed) would be priced in autarky. The home and foreign autarky asset prices of the hypothetical asset will then predict the direction of trade in the particular asset considered.

Taking the above into account, it is possible to extend the analysis to many goods and to more than two periods. As in the standard trade theory, the predictions of the law of comparative advantage are weaker for individual assets and goods, the more assets and goods there are.

The analysis has been restricted to a barter model without any money. It is clearly desirable to include the possibility of nominal assets and to analyze also the trade pattern in such assets. Extending the model to include money and other nominal assets raises several issues, though. One issue, already mentioned above, is that the appropriate gross real returns in trade equilibrium on any nominal asset considered have to be identified. We have already mentioned that the real return on nominal assets will depend on price-level risk, which in turn will depend on countries' monetary policies. For instance, difference exchange rate regimes and corresponding different monetary policies will affect the trade pattern in nominal assets and hence overall capital flows. Svensson (1987) discusses these issues and the international trade pattern for nominal assets within the context of the law of comparative advantage. Persson and Svensson (1987) examine the effect of different exchange rate regimes and corresponding exchange rate variability on capital movements within the direct approach to the determination of the trade pattern in risky assets. Another issue is that the law of comparative advantage uses the gains-from-trade theorem, which does not necessarily hold if there are domestic distortions in autarky. Hence it will be crucial for the analysis how money is modeled, more precisely whether money is modeled as having real effects and possibly being distortionary, or whether money is modeled as being neutral.

<sup>31</sup>Varian (1987) analyzes the effect on the *volume* of asset trade of differences of opinion between agents in a model with trade in Arrow-Debreu securities, using what we have called in the introductory discussion the "direct" approach. Our analysis of the effect of differences in subjective beliefs on the trade *pattern* in risky assets, using the law of comparative advantage, can hence be seen as complementary to his.

## REFERENCES

- Breeden, Douglas T., "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics*, September 1979, 7, 265-96.
- Cole, Harold, "Financial Structure and International Trade," mimeo., 1986.
- Deardoff, Alan V., "The General Validity of the Law of Comparative Advantage," *Journal of Political Economy*, October 1980, 88, 941-57.
- \_\_\_\_\_, "The General Validity of the Heckscher-Ohlin Theorem," *American Economic Review*, September 1982, 72, 693-94.
- Dixit, Avinash K. and Norman, Victor, *Theory of International Trade*, Cambridge: Nisbet and Cambridge University Press, 1980.
- \_\_\_\_\_, and Woodland, Alan D., "The Relationship between Factor Endowments and Commodity Trade," *Journal of International Economics*, November 1982, 13, 201-14.
- Dumas, Bernard, "Two-Person Dynamic Equilibrium: Trading in the Capital Market," NBER Working Paper No. 2016, 1986.
- Ethier, Wilfred J., "Higher Dimensional Issues in Trade Theory," in Ronald W. Jones and Peter B. Kenen, eds., *Handbook of International Economics*, Vol. I, Amsterdam: North-Holland, 1984, 131-84.
- \_\_\_\_\_, and Svensson, Lars E. O., "The Theorems of International Trade with Factor Mobility," *Journal of International Economics*, February 1986, 20, 21-42.
- Gordon, Robert H. and Varian, Hal R., "Taxation of Asset Income in the Presence of a World Securities Market," NBER Working Paper No. 1994, 1986.
- Grossman, Gene M. and Razin, Assaf, "International Capital Movements under Uncertainty," *Journal of Political Economy*, April 1984, 92, 286-306.
- \_\_\_\_\_, and \_\_\_\_\_, "The Pattern of Trade in a Ricardian Model with Country-Specific Uncertainty," *International Economic Review*, February 1985, 26, 193-202.
- Helpman, Elhanan, (1985a) "Comparative Advantage under Uncertainty," in H. Hesse, E. Streissler, and G. Tichy, eds., *Aussenwirtschaft bei Ungewissheit*, Tübingen: J. C. B. Mohr (Paul Siebeck), 72-88.
- \_\_\_\_\_, (1985b) "Trade Patterns under Uncertainty," Foerder Institute Working Paper No. 20-85, Tel-Aviv University.
- \_\_\_\_\_, and Razin, Assaf, *A Theory of International Trade under Uncertainty*, New York: Academic Press, 1978.
- Jones, Ronald W., "Factor Proportions and the Heckscher-Ohlin Theorem," *Review of Economic Studies*, 24, 1956, 1-10.
- Lippman, Steven A. and McCall, John J., "The Economics of Uncertainty: Selected Topics and Probabilistic Methods," in Kenneth J. Arrow and Michael D. Intriligator, eds., *Handbook of Mathematical Economics*, Vol. I, Amsterdam: North-Holland, 1981, ch. 6.
- Lloyd, Peter J. and Schweinberger, Albert G., "Trade Expenditure Functions and the Gains from Trade," mimeo., 1986.
- Lucas, Robert E., Jr., "Asset Prices in an Exchange Economy," *Econometrica*, November 1978, 46, 1429-45.
- \_\_\_\_\_, "Interest Rates and Currency Prices in a Two-Country World," *Journal of Monetary Economics*, November 1982, 10, 335-59.
- Markusen, James R. and Svensson, Lars E. O., "Trade in Goods and Factors with International Differences in Technology," *International Economic Review*, February 1985, 26, 193-202.
- Persson, Torsten and Stockman, Alan C., *Theory of International Financial Markets*, forthcoming 1988.
- \_\_\_\_\_, and Svensson, Lars E. O., "Exchange Rate Variability and Asset Trade," IIES, mimeo., 1987.
- Pomery, John, "Uncertainty in Trade Models," in Ronald W. Jones and Peter B. Kenen, eds., *Handbook of International Economics*, Vol. I, Amsterdam: North-Holland, 1984, 419-65.
- Pratt, John W., "Risk Aversion in the Small and in the Large," *Econometrica*, January-April 1964, 32, 122-36.
- Rubinstein, Mark, "The Valuation of Uncertain Income Streams and the Pricing of Options," *Bell Journal of Economics*, Autumn 1976, 7, 407-425.
- Sandmo, Agnar, "Two-Period Models of Con-

- sumption under Uncertainty: A Survey," in Jacques H. Dreze, ed., *Allocation under Uncertainty: Equilibrium and Optimality*, New York and Toronto: Wiley & Sons, 1974, ch. 2.
- Selden, Larry**, "A New Representation of Preferences over 'Certain  $\times$  Uncertain' Consumption Pairs: The 'Ordinal Certainty Equivalent' Hypothesis," *Econometrica*, September 1978, 46, 1045-60.
- \_\_\_\_\_, "An OCE Analysis of the Effect of Uncertainty on Saving under Risk Preference Independence," *Review of Economic Studies*, January 1979, 46, 73-82.
- Stockman, Alan C. and Dellas, Harris**, "International Portfolio Nondiversification and Exchange Rate Variability," University of Rochester, mimeo., 1986.
- \_\_\_\_\_, and **Hernández D., Alejandro**, "Exchange Controls, Capital Controls, and International Financial Markets," University of Rochester, mimeo., 1986.
- \_\_\_\_\_, and **Svensson, Lars E. O.**, "Capital Flows, Investment and Exchange Rates," *Journal of Monetary Economics*, March 1987, 19, 171-201.
- Stulz, René M.**, "A Model of International Asset Pricing," *Journal of Financial Economics*, December 1981, 9, 383-406.
- \_\_\_\_\_, "Currency Preferences, Purchasing Power Risks, and the Determination of Exchange Rates in an Optimizing Model," *Journal of Money, Credit, and Banking*, August 1984, 16, 302-16.
- \_\_\_\_\_, "An Equilibrium Model of Exchange Rate Determination and Asset Pricing with Non-Traded Goods," Ohio State University, mimeo., 1986.
- Svensson, Lars E. O.**, "Factor Trade and Goods Trade," *Journal of International Economics*, May 1984, 16, 365-78.
- \_\_\_\_\_, "Currency Prices, Terms of Trade and Interest Rates: A General Equilibrium Asset-Pricing Cash-in-Advance Approach," *Journal of International Economics*, February 1985, 18, 17-41.
- \_\_\_\_\_, "Trade in Nominal Assets: Monetary Policy, and Price Level and Exchange Rate Risk," IIES Seminar Paper No. 392, 1987.
- Varian, Hal R.**, "Differences in Opinion in Financial Markets," University of Michigan, mimeo., 1987.
- Woodland, Alan D.**, *International Trade and Resource Allocation*, Amsterdam: North-Holland, 1982.

# The Optimal Tariff, Production Lags, and Time Consistency

By HARVEY E. LAPAN\*

*The optimal tariff for a large country equals the reciprocal of the foreign export elasticity of supply. However, if production decisions occur before consumption decisions, the ex ante optimal tariff is not time consistent because the ex post elasticity is less than the ex ante elasticity. We show all countries are worse off if the large country cannot precommit to its ex ante optimal tariff, and that all countries can gain if the large country taxes domestic production of importables.*

It is well known that a large country can increase domestic welfare through trade restrictions. In particular, the optimal *ad valorem* tariff rate on imports equals the reciprocal of the foreign export price elasticity of supply. The operational problems with this formula are that the relevant elasticity is not a constant (as it varies along the offer curve) and it will, in general, depend upon the appropriate time perspective.

For example, assume a time lag between production and trade decisions, and that the home country is "large." From an *ex ante* perspective, the elasticity of foreign export supply depends upon both foreign supply and demand elasticities, whereas the *ex post* (given production choices) foreign export supply elasticity depends only upon preferences, and hence will be less elastic than the *ex ante* export supply schedule. Accordingly, domestic policymakers, attempting to capitalize on the home economy's greater market power, will have an incentive to set *ex post* tariffs at a higher level than they would if they could irrevocably precommit themselves to an *ex ante* tariff.

Foreign and domestic producers, aware of the tariff rule used (without precommitment) will adjust their production decisions accordingly. Consequently, both countries will be worse off in the situation, where tariffs are set *ex post*, in accord with the short-run

export supply elasticity, than they would be if tariffs were set *ex ante*. This dilemma is akin to the now familiar issue of time consistency in macro models: the inability of the government to (believably) precommit itself to a known policy can lead to a suboptimal outcome. Recent papers by Jonathan Eaton and Gene Grossman (1985) and by Robert Staiger and Guido Tabellini (1987) address the issue of time consistency in models in which tariffs are used as second-best policy instruments.<sup>1</sup>

The political reasons why the (current) government does not, or cannot, precommit itself (or *future* governments) to a predetermined tariff is not the main focus of this paper. Nevertheless, it should be clear that, in any dynamic setting, a large country will always have an *ex post* incentive to increase tariffs above the *ex ante* optimal level. Thus, unless legislation or treaties can be implemented that prohibit (future) governments from changing tariffs, the optimal *ex ante* tariff is not a credible, time-consistent solution.<sup>2</sup>

<sup>1</sup>Both papers focus on the ability of tariffs to redistribute income and both assume incomplete (insurance) markets. Eaton and Grossman report, based on numerical methods, that the optimal and time-consistent solutions are similar. The Staiger-Tabellini model indicates that the time-consistent solution leads to (more) protection.

<sup>2</sup>Reputational considerations may cause the *ex post* tariff to be set closer to the *ex ante* optimal level as the large country government attempts to persuade foreign producers that it will *not* exploit the inelasticity of the *ex post* offer curve. However, if reputation is not transferable between governments, these reputational consid-

\*Department of Economics, Iowa State University, Ames, Iowa 50011. I am indebted to Walt Enders, and two anonymous referees, for valuable comments and suggestions. Remaining errors are my responsibility.

The plan of our paper is as follows. In Section I we present the basic model, derive the time-consistent tariff, and compare it with the optimal tariff. In Section II we show that, due to the second-best nature of the time-consistent tariff, a production tax on importables in the large country can benefit all countries. We also compare the time-consistent tariff/tax equilibrium with the optimal and time-consistent solutions from Section I. The Appendix contains proofs of these propositions.

### I. The Optimal and the Time-Consistent Tariff

Our analysis utilizes the standard two-good ( $M, F$ ) trade model, with well-behaved preferences and technology. There are a large number of small, identical foreign countries that pursue free trade policies. The domestic economy is large, and would, under free trade, export  $M$  (the numeraire). All private agents act as price takers, but the domestic government utilizes commercial policy to increase domestic welfare.

The distinguishing characteristic of our model is its focus on the timing of economic decisions, which are made in the following sequence: (i) the domestic government announces the tariff rate on imports (of  $F$ ); (ii) domestic and foreign production decisions are made based upon producer prices expected to prevail when trade/consumption decisions are made; (iii) the government may revise its announced tariff prior to trade decisions; and (iv) finally, trade (and consumption) decisions are made and carried out, given the predetermined production levels and tariff rate.

Given our assumption of full information, the key assumption is not merely the production lag, but more importantly the ability of the government to revise tariffs once pro-

duction decisions are made (step (iii)). If, in step (i), the government could credibly precommit to a tariff (thereby abolishing step (iii)) then, despite the production "lag," our model, and results, would reduce to those of the standard optimal tariff literature. However, if the government can change tariffs after production decisions are made, step (i) is essentially irrelevant, and the standard optimal tariff will not be time consistent. This time inconsistency of the optimal tariff arises not only in the presence of production lags (as for agricultural markets, with spring plantings and fall harvest), but also in a dynamic model in which *ex ante* and *ex post* supply elasticities differ. In essence, precommitment means that the current government can set tariff rates for all time, and can *preclude* future governments from ever changing these rates. If the current government cannot exercise such authority, then the (dynamic) optimal tariff will not be time consistent.<sup>3</sup>

We now turn to our formal model. Let  $\bar{p}$  and  $p$ , respectively, denote the foreign, and domestic, relative (consumer) price of  $F$ . Similarly,  $\bar{p}^a$  and  $p_s^a$  denote the relative price foreign, and domestic, producers of  $F$  expect (to receive for their output) when production decisions are made. In a perfect foresight equilibrium,  $\bar{p}^a = \bar{p}$  and  $p_s^a = p$  (if no domestic production tax/subsidies are used).

The (aggregate) foreign supply and demand curves have their usual properties, with the exception that supply depends upon price expectations, whereas demand depends upon realized income and realized price. Foreign income ( $\bar{y}$ ) is given by

$$(1) \quad \bar{y}(\bar{p}, \bar{p}^a) = \bar{S}_m(\bar{p}^a) + \bar{p}\bar{S}_f(\bar{p}^a),$$

erations will be unimportant. Also, under uncertainty, if state-contingent tariffs were not feasible, precommitting to a fixed *ex ante* tariff would not be optimal. Thus, if production lags are present, and if additional information is available once production decisions are made, the choice between *ex ante* and *ex post* tariffs involves ranking second-best policies.

<sup>3</sup>I emphasize the implausibility of precommitting to the optimal tariff because several readers of an earlier version of this paper thought it important to explain why precommitment did not occur. While this would be an interesting exercise in political economy, my main point is that precommitment would matter. Moreover, since I know of no case in which a government is *irrevocably* committed to an announced tariff (or any other economic) policy, the time-consistency issue seems germane.

where  $\bar{S}_i$  denotes the foreign supply schedule for good  $i$ . Given foreign demand for  $F(\bar{Y}, \bar{p})$ , the foreign export supply curve is

$$(2) \quad \bar{X}(\bar{p}, \bar{p}^a) \equiv \bar{S}_f(\bar{p}^a) - \bar{F}(\bar{y}, \bar{p}).$$

The slope of the *ex ante* foreign export supply curve (denoted  $\bar{X}'$ ) is derived assuming  $d\bar{p} = d\bar{p}^a$  (and  $\bar{p} \equiv \bar{p}^a$ ), whereas the slope of the *ex post* export supply curve (denoted  $\partial \bar{X} / \partial \bar{p}$ ) assumes foreign production is fixed ( $d\bar{p}^a = 0$ ):

$$(3) \quad (\partial \bar{X} / \partial \bar{p}) = -[\bar{F}_p + \bar{F}_y \bar{S}_f'(\bar{p}^a)] \\ = (\bar{X}' - \bar{S}_f') \quad \text{at } \bar{p}^a = \bar{p},$$

where  $\bar{F}_p, \bar{F}_y$  denote partial differentiation of demand, and  $\bar{S}_f' (> 0)$  is the slope of the foreign supply curve.

Turning to the domestic economy, domestic production (denoted  $S_f(p_s^a)$ ) depends on anticipated producer prices, while domestic demand depends upon the consumer price and realized income. Domestic income is

$$(4) \quad y = S_m(p_s^a) + p S_f(p_s^a) \\ + (p - \bar{p})\{F(y, p) - S_f(p_s^a)\},$$

where  $F(y, p)$  denotes domestic demand for  $F$ , and the tariff revenue (the latter term in (4)) is rebated in a lump-sum fashion to households. Using the standard representative agent assumption, we let  $V(y, p)$  denote the indirect utility function for domestic agents; domestic demand is derived from this using Roy's identity.

The equilibrium trade condition is given by

$$(5) \quad F(y, p) - S_f(p_s^a) = \bar{X}(\bar{p}, \bar{p}^a).$$

The standard optimal tariff formula is derived by maximizing  $V(y, p)$ , using (1)–(5) and assuming  $\bar{p}^a \equiv \bar{p}$ ,  $p_s^a \equiv p$  (i.e., assuming perfect foresight, no production subsidies, and that tariffs are set before production decisions are made). Performing this optimi-

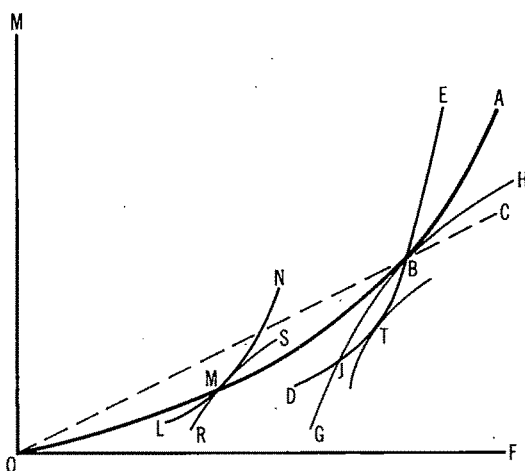


FIGURE 1. THE OPTIMAL TARIFF AND THE TIME-CONSISTENT TARIFF

zation yields the familiar result<sup>4</sup>

$$(6) \quad \{dV(y, p)/d\bar{p}\} = V_y[-\bar{X} + t\bar{X}'] = 0; \\ t \equiv (p - \bar{p}).$$

Denote this optimal *ex ante* solution by  $(p^0, \bar{p}^0)$ . As noted earlier, this solution is not time consistent if tariff rates may be changed once production decisions are made since the *ex post* foreign offer curve is less elastic than the *ex ante* curve. This time inconsistency is shown in Figure 1, where  $OMBA$  represents the *ex ante* foreign offer curve, and  $GJBH$  the domestic economy's (*ex ante*) trade indifference curve. The point  $B$  denotes the optimal trade point, with world price  $(\bar{p}^0)$  given by the slope of ray  $OBC$ , and domestic price  $(p^0)$  given by the slope of the common tangent (not shown) at  $B$ .

<sup>4</sup>Formally, we use foreign price, not the tariff, as the choice variable. If there is a unique equilibrium for each tariff, the two instruments are equivalent. However, if there are multiple equilibria for some tariffs (as may happen if the foreign offer curve is bending backward), then there is no guarantee that the "optimal" tariff will cause the "correct" equilibrium to emerge. For this case, setting world price is a superior instrument. Sufficient second-order conditions to (6) are convex domestic preferences, concave domestic technology, and a concave (*ex ante*) foreign offer curve.

Given foreign production decisions (based on  $\bar{p}^0$ ), the *ex post* foreign offer curve is less elastic, as shown by the curve *DJTBE*. While the trade point *B* is still obtainable, it is not optimal from the domestic perspective. Given foreign production, the *ex post* optimum occurs at a point like *T*, with lower foreign price (higher tariff) than at *B*.<sup>5</sup> While *T* seemingly results in higher domestic welfare than *B*, it is not a time-consistent solution since foreign producers' price expectations are incorrect. Unless the domestic government can irrevocably precommit to its *ex ante* optimal tariff, the *ex ante* optimum (*B*) is not obtainable.

The time-consistent solution requires that producers' price expectations be correct and that the *ex post* price (tariff) set by the home government be optimal, given the *ex post* offer curve. Graphically, this time-consistent equilibrium is represented in Figure 1 by point *M*, where (i) the *ex post* foreign offer curve (*LMN*) is tangent to the (relevant) domestic trade indifference curve (*RMS*), and (ii) the equilibrium occurs along the *ex ante* foreign offer curve. This time-consistent solution results in lower world price, lower foreign welfare, and lower domestic welfare.

Formally, the time-consistent, no precommitment solution is derived as follows. Using (1)–(3), equations (4) and (5) determine domestic income ( $y$ ) and price ( $p$ ) as functions of foreign price ( $\bar{p}$ ), and predetermined output levels (i.e., price expectations). For subsequent analysis, we totally differentiate (4) and (5):

$$(7) \quad dy = \left[ \{ F_p b_1 + (F + tF_p) b_2 \} / \beta \right];$$

$$t \equiv (p - \bar{p}),$$

$$(8) \quad dp = \left[ \{ -F_y(b_1) + (1 - tF_y) b_2 \} / \beta \right];$$

$$\beta \equiv [F_p + FF_y] < 0,$$

<sup>5</sup>If domestic production decisions are also made *ex ante* (based upon price expectations of  $\bar{p}^0$ ), the *ex post* trade indifference curve will not be *GJBH*. However, it will be tangent to *GJBH* at *B*, so the essential argument is unaltered.

where  $\beta$  is the slope of the domestic compensated demand for importables, and we define

$$(9) \quad b_1 \equiv \left[ -\bar{X} \cdot d\bar{p} + 0 \cdot dp^{-a} + (\bar{p} - p_s^a) S_f' dp_s^a \right],$$

$$(10) \quad b_2 \equiv \left[ (\partial \bar{X} / \partial \bar{p}) \cdot d\bar{p} + (\bar{S}_f' \cdot dp^{-a}) + (S_f' \cdot dp_s^a) \right].$$

Optimizing domestic utility ( $V(y, p)$ ), over  $\bar{p}$ , given output levels (price expectations) yields

$$(11) \quad \frac{\partial V}{\partial \bar{p}} = \left( V_y \frac{\partial y}{\partial \bar{p}} + V_p \cdot \frac{\partial p}{\partial \bar{p}} \right) = V_y \left[ -\bar{X}(\bar{p}, \bar{p}^a) + t \cdot \frac{\partial \bar{X}}{\partial \bar{p}} \right] = 0.$$

Equations (4), (5), and (11) determine the optimal foreign and domestic prices as functions of price expectations; the model is closed by assuming rational expectations and no production subsidies ( $\bar{p}^a = \bar{p}$ ,  $p_s^a = p$ ). Denote this time-consistent solution by  $(\bar{p}^c, p^c)$ .

**PROPOSITION I:** *The inability to precommit to the ex ante optimal tariff results in a lower world price and a higher domestic price of importables. It also leads to lower welfare in both countries.*

**PROOF:**

Evaluating (11) at the *ex ante* optimum  $(\bar{p}^0, p^0)$  yields

$$(12) \quad \partial V / \partial \bar{p} = V_y \left[ t^0 \left( \frac{\partial \bar{X}}{\partial \bar{p}} - \bar{X}' \right) \right] = -t^0 V_y \bar{S}_f'; \quad t^0 \equiv (p^0 - \bar{p}^0).$$

Thus, assuming  $\bar{S}_f' > 0$ ,  $\partial V / \partial \bar{p} < 0$  at  $(\bar{p}^0,$



$p^0$ ), implying  $\bar{p}^c < \bar{p}^0$ .<sup>6</sup> From (8), with  $d\bar{p}^a = d\bar{p}$ ,  $dp_s^a = dp$  (and  $p_s^a = p$ ):

$$(13) \quad (dp/d\bar{p}) = \left\{ [\bar{X}' - t\bar{S}'_f F_y] / (\beta - S'_f) \right\} < 0$$

assuming no goods are inferior. Thus,  $\bar{p}^c < \bar{p}^0$  implies  $p^c > p^0$ . Foreign welfare declines as  $\bar{p}$  falls, whereas domestic welfare must decline as the *ex ante* tariff yields the global optimum.

Since *ex post* price changes have only demand effects, it is apparent from (3) that the *ex post* foreign export supply curve will be everywhere negatively sloped if the foreign compensated price elasticity of demand is zero (no substitutability between goods). Hence, for this case, the only time-consistent solution is *autarky*! Intuitively, the time-consistent, no precommitment solution will be close to the *ex ante* optimum when the price elasticity of foreign supply is low and the demand elasticity large, whereas the equilibria will be "far apart" when foreign supply is very elastic and (compensated) demand inelastic.

Thus, the inability of the large country to irrevocably precommit to its optimal tariff leads to a decline in welfare in *both* countries. This occurs because foreign (domestic) producers, correctly anticipating a tariff above the *ex ante* optimum, respond by decreasing (increasing) production of  $F$ , thereby reducing world trade and specialization. Naturally, if the large country government could impose some cost (or treaty) on itself (or future governments) that effectively limited its ability to alter tariffs, this time-inconsistency problem could be reduced, or eliminated.

<sup>6</sup>We assume a unique *ex ante* optimum tariff, and a unique, consistent solution for the *ex post* tariff. In general, satisfaction of the *SOC* for the *ex post* tariff is neither necessary, nor sufficient, to guarantee uniqueness. Sufficient conditions for uniqueness, and the *SOC*, are: (i) convex preferences; (ii) concave technology; (iii) concave *ex post* foreign offer curve; and (iv) no inferior goods. Details are in the Appendix and an earlier version of this paper.

Failing this, it will be in the interests of the domestic government to induce foreign producers to expand output by precommitting itself to other policies that will have the net impact of raising world prices. One such policy is to limit (tax) domestic production of importables.

## II. Domestic Production Taxes and the Time-Consistent Tariff

It is well known that (with precommitment) the optimal policy for a large country entails trade restrictions, but no production taxes ( $MRS = DRT = FRT$ ). However, the time-consistent solution described in Section I results in an equilibrium in which:  $MRS = DRT > ex\ ante\ FRT$ . Intuitively, then, a policy that restricted (taxed) domestic production of importables—thereby encouraging foreign production—could increase welfare in both countries.

The timing of economic decisions, as detailed in Section I, is modified as follows: First, the government announces a price (or tax/subsidy) it will pay domestic producers, as well as a tariff rate. Next, domestic (and foreign) production decisions are made, on the basis of producer price expectations. Then, the government—assuming precommitment is not feasible—can revise the tariff rate (or production tax/subsidy). Finally, trade and consumption decisions are made. Note that time-consistency issues do *not* arise with respect to revision of the producer price (or tax) in step (iii) since, given the level of domestic production, such *ex post* changes merely redistribute domestic income, but cannot change output (or domestic "welfare").<sup>7</sup>

The *ex ante* optimal domestic producer price is determined as follows.<sup>8</sup> Equations

<sup>7</sup>If all agents are alike, revisions in the tax—given output—have no effect. If agents differ, and other forms of lump-sum transfers are not feasible, a social welfare function would be required to determine the time-consistent production tax (and tariff).

<sup>8</sup>The *ex ante* choice of domestic output, the production tax, or the producer price are equivalent instruments.

(4), (5), and (11), in conjunction with perfect foresight ( $\bar{p}^a = \bar{p}$ ), determine the (time-consistent) world and domestic prices, and domestic income, as functions of the predetermined domestic output level (or producer price,  $p_s^a$ ). Totally differentiating  $V(y, p)$  with respect to  $p_s^a$ , using (7)–(11) yields—after some simplification

$$(14) \quad dV/dp_s^a = V_y [\delta S_f' + t \bar{S}_f' (d\bar{p}/dp_s^a)] \\ = 0; \quad \delta \equiv (p - p_s^a).$$

In (14),  $\delta$  represents the (implicit) tax on domestic production of importables. Since the world price ( $\bar{p}$ ) will be inversely related to the domestic producer price,  $p_s^a$  (positively related to the production tax), it is apparent from (14) that some production tax on importables will be desirable. Denote the optimal production tax by  $\delta^*$ , and the resulting domestic and foreign prices by  $(p^*, p_s^*, \bar{p}^*)$ . Then:

**PROPOSITION II:** *Assume the large country cannot precommit to its optimal tariff, but it can control (tax or subsidize) domestic production. Then, assuming all goods are normal and the ex post foreign offer curve is concave (increasing FRT):*

(i) An optimal policy entails a tax on domestic production of importables that is less than the *ex post* tariff ( $t^* \equiv [p^* - \bar{p}^*] > \delta^* > 0$ ).

(ii) The resulting production tax, tariff equilibrium is characterized by

$$\{ \text{ex post FRT} = p^* \equiv MRS > p_s^* \\ \equiv DRT > \text{ex ante FRT} > \bar{p}^* \}.$$

(iii) Foreign nations, as well as the domestic country, gain from the production tax.

(iv) Nevertheless, the resulting equilibrium is Pareto inferior to the *ex ante* optimal tariff, with higher domestic (consumer) prices of importables, and lower world prices ( $p^* > p^c > p^0$ ;  $\bar{p}^c < \bar{p}^* < \bar{p}^0$ ).

**PROOF:**

See the Appendix.

Note that this implies that a large country which exports agricultural goods (for which production lags exist) could gain by subsidizing production of exportables, assuming tariff precommitment is not feasible.<sup>9</sup>

### III. Conclusions

We have shown that if production lags are present and tariff precommitment is not feasible, then the time-consistent tariff equilibrium is Pareto inferior to the precommitment equilibrium, and the second-best solution will include a production tax on importables. Clearly, these conclusions extend to any dynamic framework in which some inputs are committed before trade decisions are made.

An interesting extension would be to incorporate uncertainty into the model. Assuming state-contingent policies (tariffs) are not feasible, then both *ex ante* and *ex post* tariffs would be second-best instruments, as the *ex ante* tariff is set under imperfect information, whereas the *ex post* tariffs involve time-consistency problems. The relative advantage of flexibility (deferring tariff decisions) should depend upon the degree of uncertainty and the elasticity of foreign supply.

### APPENDIX: PROOF OF PROPOSITION II

Using (4)–(5), the FOC for the *ex post* tariff (11) defines

$$(A1) \quad J[\bar{p}, \bar{p}^a, p_s^a] = V_y \left[ -\bar{X} + t \frac{\partial \bar{X}}{\partial \bar{p}} \right] = 0.$$

The SOC, given  $(\bar{p}^a, p_s^a)$ , requires

$$(A2) \quad \partial J / \partial \bar{p} = (1/\beta) \left[ (\bar{X}' - \bar{S}_f')^2 - \beta \Delta \right] < 0; \\ \Delta = [2(\bar{X}' - \bar{S}_f') - t(\bar{X}'' - \bar{S}_f'' + \bar{F}_y \bar{S}_f')],$$

<sup>9</sup>Since using a production tax and a tariff is equivalent to using a production tax/subsidy and a consumption tax, Proposition II implies that the second-best policy entails a production *subsidy* (on importables) that is less than the consumption tax. It should also be clear that, in this model, the time inconsistency of the optimal tariff arises because of the “consumption-tax component” of the tariff. I am indebted to an anonymous referee for suggesting this observation.

where  $\beta$  (slope of domestic compensated demand for importables) is negative, and  $\Delta > 0$ , if and only if the *ex post* foreign offer curve is concave (i.e., increasing FRT, given output). Uniqueness of the time-consistent solution (given  $p_s^a$ ) requires

$$(A3) \quad \left[ \frac{\partial J}{\partial \bar{p}} + \frac{\partial J}{\partial \bar{p}^a} \right] = (-W/\beta) < 0;$$

$$W = [\beta \cdot K - (\bar{X}' - \bar{S}_f')(\bar{X}' - tF_y \bar{S}_f')];$$

$$(A4) \quad K = [\Delta + \bar{S}_f'(1 + t\bar{F}_y)].$$

$K > 0$  implies the *ex post* FRT increases as one moves up the *ex ante* offer curve. Assuming normality, the SOC implies uniqueness. Comparative static results are derived from (4)–(5) and (A1)–(A4):

$$(A5) \quad [d\bar{p}/dp_s^a] = [(1 - \delta F_y) S_f'(\bar{X}' - \bar{S}_f')/W] < 0;$$

$$\delta \equiv (p - p_s^a),$$

$$(A6) \quad [dp/dp_s^a] = [S_f'(1 - \delta F_y) K/W] < 0;$$

$$\text{thus } d\delta/dp_s^a < 0.$$

From (14), after substitution and simplification

$$(A7) \quad [dV/dp_s^a] = [V_y S_f'/W].$$

$$[\delta \beta K + (\bar{X}' - \bar{S}_f')(t\bar{S}_f' - \delta \bar{X}')] = 0 \Rightarrow,$$

$$(A8) \quad t^* > (t^* \bar{S}_f'/\bar{X}') > \delta^* > 0$$

since  $\bar{X}' > \bar{S}_f'$  at a consistent solution.

The *ex ante* FRT, at the consistent tariff/tax solution, is

$$(A9) \quad FRT = [\bar{p} + (\bar{X}/\bar{X}')] ]$$

$$= [\bar{p} + \{t^*(\bar{X}' - \bar{S}_f')/\bar{X}'\}]$$

$$= [p^* - (t^* \bar{S}_f'/\bar{X}')] < [p^* - \delta^*].$$

Hence, as stated,  $p^* > p_s^* > \text{ex ante FRT} > \bar{p}^*$ . From (A5) and (A6),  $\delta^* > 0 \Rightarrow p^* > p^c > p^0$  and  $\bar{p}^* > \bar{p}^c$ , which implies foreign exporters gain from the tax.

Finally, define  $\bar{p}(p_s^a)$  as the consistent solution to (11), given exogenously determined  $(p_s^a)$ , and define  $(\hat{p}_s^a)$  s.t.  $\bar{p}(\hat{p}_s^a) = \bar{p}^0$ . From Section I,  $\bar{p}(p^0) < \bar{p}^0$ ; using (A5) this implies  $\hat{p}_s^a < p^0$ . But  $p^0 = \text{ex ante FRT}$  at  $\bar{p}^0$ . Hence,  $p_s^* > \text{ex ante FRT} \Rightarrow p_s^* > \hat{p}_s^a = \bar{p}^* < \bar{p}^0$ . Thus, the production tax/tariff equilibrium is Pareto inferior to the *ex ante* optimum, completing the proof of Proposition II.

## REFERENCES

- Bhagwati, Jagdish N. and Srinivasan, T. N., *Lectures on International Trade*, Cambridge: MIT Press, 1983.
- Eaton, Jonathan and Grossman, Gene M., "Tariffs as Insurance: Optimal Commercial Policy When Domestic Markets Are Incomplete," *Canadian Journal of Economics*, May 1985, 18, 258–72.
- Graaff, J. De V., "On Optimum Tariff Structures," *Review of Economic Studies*, 1949–50, 17, 47–59.
- Karp, Larry "Optimality and Consistency in a Differential Game with Non-Renewable Resources," *Journal of Economic Dynamics and Control*, October 1984, 8, 73–97.
- Kydland, Finn and Prescott, Edward "Rules Rather Than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, 85, 473–93.
- Newbery, David M. G., "Oil Prices, Cartels and the Problem of Dynamic Inconsistency," *Economic Journal*, September 1981, 91, 617–46.
- Staiger, Robert W. and Tabellini, Guido, "Discretionary Trade Policy and Excessive Protection," *American Economic Review*, December 1987, 77, 832–37.

# Investment, Capacity Utilization, and the Real Business Cycle

By JEREMY GREENWOOD, ZVI HERCOWITZ, AND GREGORY W. HUFFMAN\*

*This paper adopts Keynes' view that shocks to the marginal efficiency of investment are important for business fluctuations, but incorporates it in a neoclassical framework with endogenous capacity utilization. Increases in the efficiency of newly produced investment goods stimulate the formation of "new" capital and more intensive utilization and accelerated depreciation of "old" capital. Theoretical and quantitative analysis suggests that the shocks and transmission mechanism studied here may be important elements of business cycles.*

In the real-business cycle models of the type developed by Finn Kydland and Edward Prescott (1982), and John Long and Charles Plosser (1983), the cycles are generated by exogenous shocks to the production function. A stylized version of the main mechanism working in these models can be described as follows. Dynamic optimizing behavior on the part of agents in the economy implies that both consumption and investment react positively to these direct shocks to output. Since the marginal productivity of labor is directly affected, employment is also procyclical. The resulting capital accumulation provides a channel of persistence, even if the technology shocks are serially uncorrelated. Hence, these productivity shocks are able to generate, from a neoclassical framework, co-movements of macroeconomic variables and persistence of fluctuations that conform to those typically observed during business cycles.

In contrast with the mechanism described above, where investment reacts to changes in output, the present paper adopts John Maynard Keynes' (1936) view that it is shocks to the marginal efficiency of investment that are important for generating out-

put fluctuations. However, these shocks are incorporated here in a neoclassical framework where the rate of capital utilization is endogenous. In the present model, a positive shock to the marginal efficiency of investment stimulates the formation of "new" capital and the more intensive utilization and accelerated depreciation of "old" capital. The main operating characteristics of the proposed model are analyzed in order to gain an understanding of the transmission mechanism of the shocks. Of particular theoretical interest are the qualitative characteristics of the pattern of co-movements and persistence effects permissible in this framework. Then, a quantitative analysis of the model is performed to assess its ability to mimic the observed pattern of postwar-U.S. business cycle fluctuations. This is carried out by constructing a parametrized version of the model for which the exact joint probability distribution of the endogenous and exogenous variables is numerically computed. Using this distribution, a set of second moments for the artificial economy's variables—reflecting their co-movements and persistence—is computed and compared with that characterizing U.S. data.

Fluctuations in investment played a key role in Keynes' view of the trade cycle. There, shifts in the marginal efficiency of investment impact on investment, aggregate demand and therefore, given the disequilibrium in the labor market, employment and output. The quintessential case of this type is when there is an increase in the

\*Department of Economics, University of Western Ontario, London N6A 5C2, Canada; Department of Economics, Tel Aviv University, Tel Aviv 69978 Israel; Department of Economics, The University of Western Ontario, London N6A 5C2 Canada. The authors thank Finn E. Kydland, Edward C. Prescott, Thomas J. Sargent, Lars E. O. Svensson, and two anonymous referees for their helpful comments.

marginal efficiency of newly produced capital that does not affect the productivity of the capital stock already on line. When a shock of this type occurs in a standard neoclassical model, employment and output also tend to rise, but the mechanism is very different. The increase in the rate of return on investment stimulates current labor effort and output through an intertemporal substitution effect on leisure. A potential problem with this mechanism, as discussed by Robert Barro and Robert King (1984), is that intertemporal substitution which induces individuals to postpone leisure, also works to cut consumption. This effect would tend to make consumption move countercyclically, which contradicts the evidence. Labor productivity would tend to move in the "wrong" direction, too. An expansion of labor effort, given the fixed supply of capital in the short run, causes labor's productivity to decline.

In contrast to the intertemporal substitution effect mentioned above, the transmission mechanism of the investment shocks works in the present model through the optimal utilization of capital and its positive effect on the marginal productivity of labor. As will be seen, an important aspect of such a change in labor productivity is that it creates *intratemporal* substitution, away from leisure and toward consumption, generating procyclical effects on consumption and labor effort. Additionally, average labor productivity responds procyclically to these shocks.

To sharpen the distinction between this and the real business cycle models with direct shocks to the production function, no shifts of the latter type are included. Therefore, given the quantities of capital and labor input, current productivity shifts are endogenous in this framework. The shocks to investment are modeled as current technological changes that affect the productivity of new capital goods only, leaving unchanged the productivity of the existing capital. Because of a time-to-build delay, only the productivity of future capital is affected. This type of technological change may be more realistic than the current shock to productivity. Important technological improvements of new productive capital seem to occur quite often. As will be discussed, it is crucial

for this model that the new technology does not affect directly the productivity of the existing capital stock.

A description of the environment characterizing the economy under study is given in Section I. In Section II the representative agent's optimization problem is cast. The theoretical investigation of the model is carried out in Section III, and the quantitative analysis in Section IV. Finally, concluding remarks are offered in Section V.

### I. The Economic Environment

Consider a perfectly competitive closed economy populated by a very large number of identical households and identical firms. Aggregate output is given by the following constant-returns-to-scale production function which differs from the standard neoclassical specification solely by the inclusion of a variable rate of capital utilization

$$(1) \quad y_t = F(k_t, h_t, l_t),$$

where  $y_t$  is the output of the single good in period  $t$ ,  $k_t$  is the capital stock (see below for a discussion about its units) at the beginning of period  $t$ ,  $h_t$  is an index of the period- $t$  utilization rate of  $k_t$ , and  $l_t$  is labor input in this period. The variable  $h_t$ —which for a given capital stock determines the flow of capital services  $k_t h_t$ —represents the intensity of the use of capital, that is, the speed of operation or the number of hours per period the capital is used. An alternative interpretation of  $h_t$  is that while  $l_t$  represents the total labor employed,  $h_t$  reflects the portion of it used directly in production, with the remainder being involved in maintenance activities. The nonnegative, constant-returns-to-scale function  $F$  satisfies  $F_1, F_2 > 0$ ,  $F_{11}, F_{22} < 0$ , and  $F_{11}F_{22} - F_{12}^2 = 0$ . A consequence of the constant-returns-to-scale assumption is that  $F_{12} > 0$ , which implies capital and labor services are complements in the Edgeworth-Pareto sense. This feature provides a positive link between capital utilization and labor productivity.<sup>1</sup>

<sup>1</sup>A special case would be one of fixed proportions where, for a given  $k_t$ ,  $h_t$  and  $l_t$  should move together.

The capital utilization decision involves Keynes' notion of "user cost." That is, a higher utilization rate causes a faster depreciation of the capital stock, either because wear and tear increase with use or because less time can be devoted to maintenance.<sup>2</sup> As in the work of Paul Taubman and Maurice Wilkinson, 1970; Guillermo Calvo, 1975; John Merrick, 1984; and Zvi Hercowitz, 1986, this effect is modeled in the evolution of the capital stock as

$$(2) \quad k_{t+1} = k_t[1 - \delta(h_t)] + i_t(1 + \varepsilon_t),$$

where the nonnegative depreciation function  $\delta$  satisfies  $0 \leq \delta \leq 1$ ,  $\delta' > 0$ ,  $\delta'' > 0$ . Gross investment, as corresponding to the national income accounts, is  $i_t$ . Its contribution to the production capacity in  $t+1$ , however, depends on the technological shift factor  $\varepsilon_t$ , affecting the productivity of the new capital goods. The productivity of the already installed capital stock  $k_t$  is not directly affected by the new technology. Correspondingly,  $k_{t+1}$  is a measure of the future capital stock in productivity units. (Similarly  $k_t$  would include past technological changes.)

Note that this technological disturbance is very different from the usual technological shock, attached to the production function, used in the real business cycle models. By substituting  $k_{t+1}$  into the production function corresponding to  $t+1$ , it becomes clear that  $\varepsilon_t$  works as a shift in the marginal efficiency of capital produced in period  $t$  which comes on line in  $t+1$ . The length of the basic period, which corresponds to the time-to-build, is thought of as nontrivial, say one year (see the discussion in Kydland and Prescott, 1982).

The value of  $\varepsilon_t$ , which is realized at the beginning of period  $t$ , is generated from the stationary Markov distribution function  $\Phi(\varepsilon_t | \varepsilon_{t-1})$  defined on the domain  $Q = [\underline{\varepsilon}, \bar{\varepsilon}]$ .

<sup>2</sup> Keynes said: "User cost constitutes the link between the present and the future. For in deciding his scale of production an entrepreneur has to exercise a choice between using up his equipment now or preserving it to be used later on..." (1936, pp. 69-70) [quoted also by Taubman and Wilkinson, 1970].

The assumption that  $\varepsilon_t$  is stationary implies in the present framework, that the equilibrium capital stock will also be stationary.

The representative household in this economy maximizes expected lifetime utility as given by

$$(3) \quad E_0 \left[ \sum_{t=0}^{\infty} \beta^t \underline{U}(c_t, l_t) \right] \quad 0 < \beta < 1,$$

where  $c_t$  and  $l_t$  are the period- $t$  flows of consumption and labor effort, and  $\beta$  is the discount factor.

The specific form of  $\underline{U}$  adopted is

$$\underline{U}(c_t, l_t) = U(c_t - G(l_t)),$$

with  $U' > 0$ ,  $U'' < 0$ ,  $G' > 0$ , and  $G'' > 0$ , and where  $U$  is assumed to be bounded from above. This utility function satisfies the standard properties  $\underline{U}_1 > 0$ ,  $\underline{U}_2 < 0$ ,  $\underline{U}_{11}$ ,  $\underline{U}_{22} < 0$ ,  $\underline{U}_{11}\underline{U}_{22} - \underline{U}_{12}^2 > 0$ , and it implies that the marginal rate of substitution between consumption and labor effort depends on the latter only:

$$-\frac{\underline{U}_2(c_t, l_t)}{\underline{U}_1(c_t, l_t)} = G'(l_t).$$

That is, labor effort is determined independently of the intertemporal consumption-savings choice, which is very convenient in obtaining results from the model. As a consequence, the intertemporal substitution effect on labor effort, a central ingredient in many macroeconomic models, is eliminated. Rather than being a drawback, this implication of the utility function has the advantage of emphasizing the alternative transmission of investment shocks being studied here. When analyzing fluctuations in labor effort, this framework stresses shifts in the productivity of labor brought about by changes in the optimal rate of capacity utilization, as opposed to intertemporal substitution effects stressed by others.

The description of the setup is completed by the resource constraint

$$(4) \quad y_t = c_t + i_t.$$

## II. The Representative Agent's Optimization Problem

The decision making of consumer-workers and firms in competitive equilibrium can be summarized by the outcome of the following "representative" agent's dynamic-programming problem

$$(5) \quad V(k_t; \varepsilon_t) = \max_{(c_t, k_{t+1}, h_t, l_t)} \left[ \underline{U}(c_t, l_t) + \beta \int_Q V(k_{t+1}; \varepsilon_{t+1}) d\Phi(\varepsilon_{t+1} | \varepsilon_t) \right],$$

subject to

$$(6) \quad c_t = F(k_t, h_t, l_t) - \frac{k_{t+1}}{1 + \varepsilon_t} + \frac{k_t}{1 + \varepsilon_t} \times [1 - \delta(h_t)],$$

where the transition equation (6) is obtained by substituting the production function (1) and the capital evolution equation (2) into the resource constraint (4).<sup>3,4</sup> It can be established that the value function  $V(\cdot; \cdot)$  exists, is unique, increasing, concave, and differentiable in its first argument (see

<sup>3</sup>The functions  $F(\cdot)$ ,  $\delta(\cdot)$ ,  $U(\cdot)$ , and  $G(\cdot)$  are all assumed to be twice continuously differentiable.

<sup>4</sup>Note that the capacity utilization variable,  $h$ , could be eliminated from the above programming problem by utilizing the production function defined by

$$\phi(k, l, \varepsilon) = \max_h \left[ F(kh, l) + \frac{k(1 - \delta(h))}{(1 + \varepsilon)} \right].$$

It is easy to establish that  $\phi(\cdot)$  is well-behaved in the usual sense of being jointly concave in  $k$  and  $l$ , etc. Of interest is the fact that while a technological improvement increases the marginal product of labor, it decreases that of capital. Thus,  $\varepsilon$  does not operate here in the manner of standard Hicks or Harrod-neutral technological shock. Using this production function, equation (6) can be rewritten as

$$c_t = \phi(k_t, l_t, \varepsilon_t) - \frac{k_{t+1}}{(1 + \varepsilon_t)},$$

which does not involve  $h_t$ .

Robert Lucas, Edward Prescott, and Nancy Stokey, 1985).

The solution to the above programming problem is characterized by the following three efficiency conditions—in addition to (6)

$$(7) \quad U'(c_t - G(l_t))/(1 + \varepsilon_t) = \beta \int_Q V_1(k_{t+1}; \varepsilon_{t+1}) d\Phi(\varepsilon_{t+1} | \varepsilon_t) = \beta \int_Q U'(c_{t+1} - G(l_{t+1})) \times [F_1(k_{t+1}h_{t+1}, l_{t+1})h_{t+1} + (1 - \delta(h_{t+1}))(1 + \varepsilon_{t+1})] \times d\Phi(\varepsilon_{t+1} | \varepsilon_t).$$

$$(8) \quad F_1(k_t, h_t, l_t) = \delta'(h_t)/(1 + \varepsilon_t).$$

$$(9) \quad F_2(k_t, h_t, l_t) = G'(l_t).$$

The first equation (7) is a standard optimality condition governing investment. The left-hand side of this equation represents the loss in current utility which is realized when an extra unit of current investment is undertaken. The right-hand side portrays the discounted expected future utility obtained from an extra unit of investment today. Note that an increase in the investment technological shift factor,  $(1 + \varepsilon_t)$ , reduces the utility cost of an extra unit of capital accumulation in this period. This occurs because a given increase in expected future output can now be obtained with a lower amount of current investment.

The next equation (8) characterizes efficient capital utilization. It states that capital should be utilized at the rate,  $h_t$ , which sets the marginal benefit of capital services equal to the marginal user cost. The marginal user cost of capital is made up of two components. Specifically,  $\delta'(h_t)$  represents the marginal cost in terms of increased current depreciation from utilizing capital at a higher rate, while  $1/(1 + \varepsilon_t)$  is the current replacement cost of old in terms of new capital.

Finally, equation (9) sets the marginal product of labor equal to the marginal disutility of working, measured in terms of consumption. Again, given the form of the utility function adopted, the latter depends only upon current labor effort, and thus is determined independently of the agent's intertemporal consumption-savings decision. The advantage of this characteristic is that the system of equations (6)–(9) is recursive in the sense that (8) and (9) jointly determine  $h_t$  and  $l_t$ , while then given these solutions equation (7)—in conjunction with (6)—determines the intertemporal allocation, which amounts here to specifying values for  $k_{t+1}$  and  $c_t$ .

### III. Qualitative Analysis of the Model

An analysis of the effect of the technology shift  $\varepsilon_t$  governing the marginal efficiency of investment, on output, hours worked, capacity utilization, productivity, investment, and consumption will now be undertaken. For simplicity, the discussion in this section is carried out under the assumption that the disturbances are purely temporary—that is, independently distributed over time, so that  $\Phi(\varepsilon_{t+1}|\varepsilon_t) = \Phi(\varepsilon_{t+1})$ . This serves two purposes: first, it allows for clear results to be obtained, and second, it emphasizes the main characteristics of the model's propagation mechanism.

#### A. Impact Effect of Investment Shocks

The endogenous capital utilization is central to the model's ability to generate positive co-movement of investment, productivity, and consumption. In a neoclassical model of the present type but with constant capital utilization (and a general utility function), shocks to the productivity of investment would have different effects. A positive shock, for example, would tend to generate intertemporal substitution, away from current consumption and toward current investment and future consumption. The higher return on currently available resources would, at the same time, operate to persuade individuals to postpone leisure. Additionally, the resulting expansion in labor effort and out-

put would lead to a decline in labor productivity, given the fixed stock of capital in place. Thus, investment and output would move inversely with consumption and productivity in response to these shocks. (A formal discussion of the above is provided in the Appendix.)

Given the structure of the optimality conditions (7)–(9), the effect on the variables,  $h_t$ ,  $l_t$ ,  $y_t$ , and productivity, can be calculated from (8) and (9) only. Performing the standard comparative statics exercise on (8) and (9) yields

$$(10) \quad \frac{dh_t}{d\varepsilon_t} = -\delta'(t)[F_{22}(t) - G''(t)]/$$

$$[(1 + \varepsilon_t)^2 \Omega(t)] > 0,$$

$$(11) \quad \frac{dl_t}{d\varepsilon_t} = F_{12}(t)k_t\delta'(t)/$$

$$[(1 + \varepsilon_t)^2 \Omega(t)] > 0,$$

with

$$\Omega(t) \equiv -F_{11}(t)k_tG''(t) - \delta''(t) \\ \times [F_{22}(t) - G''(t)]/(1 + \varepsilon_t) > 0,$$

where the sign restriction follows from the concavity of  $F(\cdot)$ , and the convexity of  $\delta(\cdot)$  and  $G(\cdot)$ .

The interpretation of these results is that  $\varepsilon_t$  reduces the cost of capital utilization and hence induces a higher  $h_t$ . Since  $F_{12} > 0$ , labor's marginal productivity increases, resulting in a higher level of employment.<sup>5</sup>

<sup>5</sup>A labor market interpretation of these results is the following. Letting  $w_t$  represent the period- $t$  real wage, equilibrium in the labor market for this period can be characterized by the condition  $l^d(k_t, h_t, w_t) = l^s(w_t)$ , where the labor demand function  $l^d(t)$  solves the equation  $F_2(k_t, h_t, l^d(k_t, h_t, w_t)) = w_t$  and the labor supply function,  $l^s(t)$ , is given by  $l^s = G'^{-1}(w_t)$ . Clearly, increased capacity utilization induces a positive shift in labor demand which necessitates equilibrating increases in both the real wage and labor supply. By contrast, in the conventional model with time-separable preferences and with constant capacity utilization—see Barro and King, 1984—the period- $t$  labor market-clearing condition would be represented by an equation of the form  $l_t^d(k_t, w_t) = l_t^s(w_t, r_t, a_t)$ , where  $r_t$  represents the known period- $t$  real interest—on one-period bonds maturing in period  $t+1$ —and  $a_t$  agents' real wealth net of labor



Given that  $k_t$  is predetermined, (10) and (11) immediately imply a positive output effect. The increase in capital utilization implies that labor productivity also rises. Using (10) and (11) it is easy to establish that the marginal product of labor  $F_2(k, h_t, l_t)$ , moves upward. Specifically, one finds

$$\frac{dF_2(t)}{d\epsilon_t} = [F_{12}(t)k_t\delta'(t)G''(t)] / [(1+\epsilon_t)^2\Omega(t)] > 0.$$

Given the constant-returns-to-scale assumption, the average product of labor,  $F(k, h_t, l_t)/l_t$ , also must rise.<sup>6</sup>

The impact effects of the investment shock,  $\epsilon_t$ , on next period's capital stock,  $k_{t+1}$ , and current consumption,  $c_t$ , are deduced by displacing the system of equations (6) and (7) while making use of the first-order conditions (8) and (9). The resulting expressions are

$$(12) \quad \frac{dk_{t+1}}{d\epsilon_t} = \frac{-U'(t)}{\left[ U''(t) + \beta(1+\epsilon_t)^2 \int_Q V_{11}(t+1) d\Phi \right]} + i_t \frac{U''(t)}{\left[ U''(t) + \beta(1+\epsilon_t)^2 \int_Q V_{11}(t+1) d\Phi \right]} > 0,$$

income in this period. Here the impact of a shift in the technology factor  $\epsilon_t$  affects the current level of employment and the real wage via the intertemporal substitution effect on labor supply exerted by the induced shift in the real interest rate,  $r_t$ . (It should perhaps be emphasized that  $w_t$ ,  $r_t$ , and  $a_t$  in equilibrium will all be functions of the current state of the world.)

<sup>6</sup>This can be shown as follows. The marginal product increases if and only if the capital to labor services ratio,  $k_t h_t / l_t$ , also increases, since  $F_2(k, h_t, l_t) = F(k, h_t / l_t, 1) - (k, h_t / l_t) F_1(k, h_t / l_t, 1)$ . This is relevant since the average product  $F(k, h_t, l_t) / l_t$  can be expressed as a strictly increasing function of  $k_t h_t / l_t$ :  $F(k, h_t, l_t) / l_t = F(k, h_t / l_t, 1)$ . Therefore, average productivity also moves procyclically.

and

$$(13) \quad \frac{dc_t}{d\epsilon_t} = \frac{F_2(t)k_t F_{12}(t)\delta'(t)}{(1+\epsilon_t)^2\Omega(t)} + \frac{U'(t)/(1+\epsilon_t)}{\left[ U''(t) + \beta(1+\epsilon_t)^2 \int_Q V_{11}(t+1) d\Phi \right]} + i_t \frac{\beta(1+\epsilon_t) \int_Q V_{11}(t+1) d\Phi}{\left[ U''(t) + \beta(1+\epsilon_t)^2 \int_Q V_{11}(t+1) d\Phi \right]} \geq 0.$$

Note that these formulas presume that  $V(\cdot)$  is a twice continuously differentiable concave function in  $k$ , whereas actually it can only be shown that  $V(\cdot)$  is continuously differentiable and concave.<sup>7</sup> An argument analogous to that used by Thomas Sargent (1980) can be used to show, though, that since  $V(\cdot)$  is concave the sign restrictions in equations (12) and (13) continue to hold if these expressions are suitably reinterpreted as representing finite differences instead of derivatives.

As can be seen from (12), the technology shock,  $\epsilon_t$ , has two effects on the period- $t+1$  capital stock. The first term illustrates the positive substitution effect that an increase in the productivity of newly produced capital has on the period- $t+1$  capital stock. The second term represents the income effect associated with the shock, which is positive if  $i_t > 0$ . A given desired level for next period's capital stock can now be obtained with a lower level of current investment. Consumption-smoothing agents will utilize part of this savings in current resource utilization to increase the future stock of capital.

Current consumption is affected in three ways by a movement in  $\epsilon_t$  (compare (13)). The second term, which is negative, il-

<sup>7</sup>The agent's choice set is convex since the production function  $\phi(k, l, \epsilon)$  (see fn. 4) is concave. Therefore, standard dynamic programming arguments establish the concavity of the value function.

illustrates the *intertemporal* substitution effect associated with the improved productivity of newly produced capital. The increase in the rate of return on current investment operates to dissuade consumption and promote capital accumulation. The income effect associated with this technological change, which was explained above, works to raise current consumption and is represented by the third term. The standard macroeconomic presumption is that the intertemporal substitution effect generated by such technological shift will dominate the income effect, a situation ensured if the initial level of investment is small enough. The new element that the present model introduces is the first term, which has to do with the *intratemporal* margin of substitution between consumption and leisure. This effect may be interpreted as follows. Since  $F_{12} > 0$  a higher utilization rate increases the marginal productivity of labor, which represents the opportunity cost of current leisure in terms of consumption. This generates a substitution effect, away from leisure and toward consumption. Hence, the present model provides a channel by which both consumption and investment can possibly react procyclically.

Finally, the impact effect on gross investment,  $i_t$ , is given by

$$(14) \quad \frac{di_t}{d\epsilon_t} = \left( \frac{1}{1 + \epsilon_t} \right) \times \left[ \frac{dk_{t+1}}{d\epsilon_t} - i_t + \delta'(t)k_t \frac{dh_t}{d\epsilon_t} \right].$$

The first two terms loosely represent opposite "substitution" and "income" type effects. If the initial  $i_t$  is relatively small, the substitution effect will clearly dominate. Here there is another positive effect on gross investment coming from the additional depreciation term  $\delta'(t)k_t dh_t/d\epsilon_t$ .

The results obtained so far depend crucially upon the assumption that the technological shift pertains only to newly produced capital goods. Suppose alternatively that it applies both to newly produced,  $i_t$ , and existing capital,  $[1 - \delta(h_t)]k_t$ . Then, equation

(2) governing the evolution of capital becomes

$$k_{t+1} = k_t [1 - \delta(h_t)](1 + \epsilon_t) + i_t(1 + \epsilon_t),$$

and the transition equation (6) now becomes

$$c_t = F(k_t, h_t, l_t) - \frac{k_{t+1}}{1 + \epsilon_t} + k_t [1 - \delta(h_t)].$$

While the *form* of the efficiency conditions (7) and (9) characterizing the optimal choices for  $k_{t+1}$  and  $l_t$  remain unchanged, equation (8) specifying the optimal level for  $h_t$  is significantly altered to

$$F_1(k_t, h_t, l_t) = \delta'(h_t).$$

Since the productivity term,  $\epsilon_t$ , no longer enters the system of equations (8) and (9) now, the positive effects of a technological shift on  $h_t$ ,  $l_t$ ,  $y_t$ , and productivity, in addition to the procyclical effect on consumption, are all lost. This result obtains since it no longer pays to depreciate "off" old capital through higher levels of utilization.

### B. Dynamic Effects of Investment Shocks

Under the assumption made that  $\epsilon_t$  is serially uncorrelated, the only channel through which persistence can be generated is  $k_{t+1}$ . In the standard paradigm, a higher  $k_{t+1}$  implies more capital services, which directly tends to prolong the initial effects. In the present model, where the utilization is endogenous, higher capital does not obviously mean higher capital services. Whether there are prolonged output effects depends on how  $k_{t+1}$  affects decisions at  $t+1$ , and in particularly capacity utilization. Since the state of the world has yet to materialize, the goal here is to discern how the increase in current capital accumulation, induced by the technology shock, impacts on the means of the distributions of future endogenous variables.

From the optimality conditions (8) and (9) corresponding to period- $t+1$ , it follows that for *any given* realization of the period- $t+1$

technology shock

$$(15) \quad \frac{dh_{t+1}}{dk_{t+1}} = \left\{ -F_{11}(t+1)h_{t+1} \right. \\ \times [F_{22}(t+1) - G''(t+1)] \\ \left. + F_{12}(t+1)^2 h_{t+1} \right\} / \Omega(t+1) \\ < 0,$$

and

$$(16) \quad \frac{dl_{t+1}}{dk_{t+1}} = [\delta''(t+1)F_{12}(t+1)h_{t+1}] / \\ (1 + \varepsilon_{t+1})\Omega(t+1) > 0,$$

with the signs of the above expressions following from the facts that  $\Omega$ ,  $F_{12} > 0$ , and  $F$  is concave. The optimal rate of utilization declines since the higher  $k_{t+1}$  reduces the marginal productivity of capital services *ceteris paribus*. However, this is only a partial offsetting. The optimal flow of capital services  $k_{t+1}h_{t+1}$  increases

$$(17) \quad \frac{d(k_{t+1}h_{t+1})}{dk_{t+1}} \\ = h_{t+1} + k_{t+1} \frac{dh_{t+1}}{dk_{t+1}} \\ = -h_{t+1}\delta''(t+1) \\ \times [F_{22}(t+1) - G''(t+1)] / \\ [(1 + \varepsilon_{t+1})\Omega(t+1)] \\ > 0.$$

From (16) and (17) it follows that  $dy_{t+1}/dk_{t+1} > 0$ , for any given value of  $\varepsilon_{t+1}$ . The effects will persist also beyond  $t+1$  because from equation (7) and the first-order conditions (8) and (9) updated one period it transpires that

$$(18) \quad \frac{dk_{t+2}}{dk_{t+1}} \\ = \frac{U''(t+1)\{(1 + \varepsilon_{t+1})F_1(t+1)h_{t+1} + [1 - \delta(t+1)]\}}{[U''(t+1) + \beta(1 + \varepsilon_{t+1})^2 \int V_{11}(t+2) d\Phi]} \\ > 0.$$

Finally, to see how the expected values of the period- $t+1$  endogenous variables are affected by a period- $t$  technology shock note that these variables are functions of the period- $t+1$  state of the world—indeed this fact has been already repeatedly used—so that one can write policy functions of the form  $x_{t+1} = x(k_{t+1}, \varepsilon_{t+1})$  for  $x = h, l, hk$ , and  $y$ . It immediately transpires that  $E_t[x_{t+1}] = \int_Q x(k_{t+1}, \varepsilon_{t+1}) d\Phi(\varepsilon_{t+1})$  from which it follows that  $dE_t[x_{t+1}]/dk_{t+1} = E_t[dx_{t+1}/dk_{t+1}]$ . Thus, by taking expected values of the above expressions it obtains that period- $t+1$  labor supply, capital services, output, and period- $t+2$  capital stock all rise in expected value, while expected- $t+1$  capital utilization falls. The effects persist in similar fashion into the future periods,  $t+2, \dots$

The preceding analysis of the impact and dynamic effects of investment shocks was carried out under the assumption that  $\varepsilon_t$  is serially uncorrelated. The results about the impact effects on  $h_t$ ,  $l_t$ , and hence on current output are unchanged if  $\varepsilon_t$  is not serially independent. Note that equations (8) and (9), determining  $h_t$  and  $l_t$ , involve only  $\varepsilon_t$ , and not its future values, and hence, for these variables the serial correlation properties of  $\varepsilon_t$  are not relevant. However, it turns out that if  $\varepsilon_t$  is serially correlated, then it is not possible to sign  $dk_{t+1}/d\varepsilon_t$  unambiguously anymore.

The analysis carried out so far has shown that, theoretically, the model has potential for explaining the characteristics of business cycles. It suggests that it may be fruitful to incorporate jointly the rather Keynesian notions of shocks to the marginal efficiency of investment and variable capacity utilization into real-business cycle models. While the results obtained so far are illustrative, little light on their practical importance has been shed. Hence, a quantitative analysis of the model is now turned to with the purpose of evaluating its empirical relevance.

#### IV. Quantitative Analysis of the Model

In this section the model is suitably parameterized, calibrated, numerically solved, and evaluated. The exact nature of

the experiment being proposed is this: First, a parametric representation of the model is obtained. Second, values for the various taste and technology parameters are chosen using information from either the literature or U.S. data. Third, by varying the parameters governing the stochastic structure of the sample economy, the model is calibrated so that it yields the same standard deviation and first-order autocorrelation coefficient for output as is displayed by U.S. data. Hence, the idea is to calibrate the model to mimic the behavior of output only, both in the terms of the volatility and persistence of its fluctuations. Fourth, the model is evaluated by comparing the generated standard deviations, serial correlations, and cross correlations with output of the other variables (consumption, investment, hours, and productivity) with the corresponding statistics in the U.S. data. It should be mentioned that in undertaking the above experiment the exact stationary joint distribution for the sample economy's state variables—the capital stock and the technology shock—is numerically computed so that the (population) second moments in question can be calculated.

#### A. Sample Economy and Solution Technique

To begin with, let tastes and technology be specified in the following way:

$$\underline{U}(c, l) = \frac{1}{1-\gamma} \left[ \left( c - \frac{l^{1+\theta}}{1+\theta} \right)^{1-\gamma} - 1 \right],$$

$$F(kh, l) = (kh)^{\alpha} l^{1-\alpha},$$

$$\text{and } \delta(h) = \frac{1}{\omega} h^{\omega},$$

where  $\gamma, \theta > 0$ ,  $0 < \alpha < 1$ ,  $\omega > 1$ . Next, suppose that the stochastic structure of the environment is described by a two-state Markov process. Specifically, in any given period the technology shock,  $\varepsilon$ , is assumed to have a value lying in the time-invariant two-point set

$$E = \{e^{\xi_1} - 1, e^{\xi_2} - 1\}.$$

The distribution function governing the

drawing of a value for next period's technology shock,  $\varepsilon'$ , conditional upon a realized value for the current shock,  $\varepsilon$ , is defined by

$$\text{prob}[\varepsilon' = e^{\xi_s} - 1 | \varepsilon = e^{\xi_r} - 1] \equiv \pi_{rs},$$

where

$$0 \leq \pi_{rs} \leq 1, \quad \text{and} \quad \pi_{r1} + \pi_{r2} = 1, \\ \text{for } r, s = 1, 2.$$

The long-run (or unconditional) distribution function for the technology shock, associated with the above conditional distribution specifying the one-step transition probabilities between states, is given by<sup>8</sup>

$$(19) \quad \text{prob}[\varepsilon = e^{\xi_s} - 1] \equiv \phi_s^* \\ = \frac{\pi_{rs}}{\pi_{12} + \pi_{21}}$$

for  $r, s = 1, 2$  and  $r \neq s$ .

Finally, it will be assumed that the capital stock in each period is constrained to be an element of the finite time-invariant set  $K = \{k_1, \dots, k_n\}$ . Thus, the state space  $K \times E$  for this economy will be discrete; a similar discretization procedure has been utilized by Sargent (1980). A discussion about the assignment of values for the model's parameters— $\gamma$ ,  $\theta$ ,  $\beta$ ,  $\alpha$ ,  $\omega$ , the  $\pi$ 's, the  $\xi$ 's, and the  $n$  elements of the set  $K$ —will be postponed until later.

The representative agent's dynamic-programming problem for the above setting can be expressed as

$$(20) \quad V(k_i; \xi_r) \\ = \max_{k' \in K} \left\{ \frac{1}{1-\gamma} \left[ \left( c - \frac{\hat{l}^{1+\theta}}{1+\theta} \right)^{1-\gamma} - 1 \right] \right. \\ \left. + \beta \sum_{s=1}^2 \pi_{rs} V(k'; \xi_s) \right\}$$

$$\text{s.t. } c = (k_i \hat{h})^{\alpha} \hat{l}^{1-\alpha} \\ - k' e^{-\xi_r} + k_i \left( 1 - \frac{\hat{h}^{\omega}}{\omega} \right) e^{-\xi_r},$$

<sup>8</sup>See fn. 10 for a discussion of how these probabilities are determined.

where

$$\hat{h}, \hat{l} = \operatorname{argmax} \left[ (k_i h)^{\alpha} l^{1-\alpha} - k_i \left( 1 - \frac{h^{\omega}}{\omega} \right) e^{-\xi_r} - \frac{l^{1+\theta}}{1+\theta} \right].$$

Hence,  $h$  and  $l$  can be solved first in terms of  $k$  and  $\varepsilon$ , reducing the dynamic problem to one of choosing only  $k'$ , the next period's capital stock.

The above problem can be solved numerically using standard algorithms discussed by Dimitri Bertsekas (1976). These algorithms are based on the fact that functional equations such as (20) describe contraction mappings. This implies that iterative procedures, such as the one described below, can be used to solve numerically for the value function,  $V(\cdot)$ , over each of the  $2n$  possible points in the state space,  $K \times E$ ; that is, for a value of  $V(k_i, \xi_r)$  for each possible combination of  $k_i$  and  $\xi_r$ . To begin with, an initial guess for the value function,  $V^0(\cdot)$ , is made, say  $V^0(k_i, \xi_r) = 0$  for each  $i = 1, \dots, n$  and  $r = 1, 2$ . This initial guess for  $V(\cdot)$  is used on the right-hand side of (20) and the optimized value of the maximand, which represents the left-hand side of the equation, is used as a revised guess, or  $V^1(\cdot)$ . Then,  $V^1(\cdot)$  is entered into the right-hand side of (20) and the whole procedure is repeated until the agents' decision rules—here  $k' = k'(k, \varepsilon)$ —have converged. Convergence of the decision rules is generally faster than for the value function, the latter whose solution is generally of no intrinsic interest for the problem being analyzed (see Bertsekas, 1976, p. 245).<sup>9</sup>

From the solution to the above-programming problem the long-run or asymptotic joint distribution function of the technology shock and the equilibrium capital stock can be obtained. To see how this is done, note that the solution for next period's capital stock,  $k'$ , is such that given an initial capital

stock,  $k_i$ , and a value for the technology shock,  $\xi_r$ , a unique value for  $k' = k'(k_i, \xi_r) \in K$  will be chosen. Thus the probability  $\operatorname{prob}[k' = k_j | k = k_i, \xi = \xi_r]$  will equal one for some  $j \in \{1, \dots, n\}$  and will be zero for the rest, where trivially then

$$\sum_{j=1}^n \operatorname{prob}[k' = k_j | k = k_i, \xi = \xi_r] = 1$$

for all  $(k, \xi) \in K \times E$ .

Accordingly, the transition probability  $p_{ir,js}$  of moving from the state characterized by capital stock  $k_i$  and shock  $\xi_r$  to the one represented by  $k_j$  and  $\xi_s$  can be expressed as

$$p_{ir,js} = \operatorname{prob}[k' = k_j | k = k_i, \xi = \xi_r] \pi_{rs} \\ \forall i, j = 1, \dots, n. \\ \forall r, s = 1, 2.$$

Next the  $2n \times 2n$  transition matrix  $P$  with elements  $p_{ir,js}$  is formed. Now suppose one is arbitrarily given some initial probability distribution over the permissible values of the capital stock and technology shock. Such an initial probability distribution will be represented by the  $1 \times 2n$  vector  $\rho^0$ , specifying a probability  $\rho_{ir}^0$  that the initial capital stock/technology shock combination is  $(k_i, \xi_r)$  for each  $i$  and  $r$  pair. The probability distribution,  $\rho^1$ , governing next period's capital stock/technology shock combination is simply given by the mapping  $\rho^1 = \rho^0 P$ . Assuming this finite state Markov chain model possesses a unique asymptotic joint distribution for the capital stock and technology shock, it can then be shown that iterations on this mapping must converge to a unique fixed point,  $\rho^*$ , which satisfies  $\rho^* = \rho^* P$ .<sup>10</sup> This is true for all initial distributions  $\rho^0$ .

<sup>9</sup>In practice this iterative scheme can be accelerated using variations on the algorithm which are outlined in Bertsekas (1976).

<sup>10</sup>Similarly, one could define  $\Pi$  as the  $2 \times 2$  transition matrix associated with the technology shock whose elements are the  $\pi_{rs}$ 's. The unique steady-state distribution function associated with the technology shock,  $\phi^*$ , therefore solves the equation  $\phi^* = \phi^* \Pi$ . It is easy to deduce that the solution to this expression is given by formula (19).

Once the stationary joint probability distribution function for the capital stock and technology shock is known, it is easy to calculate various population moments of interest for the model. Note that all of the model's endogenous variables can be uniquely expressed as functions of the current capital stock and the state of technology, so that one may write  $x = x(k, \xi)$  for  $x = c, k', h, l, y$ , etc. Thus, for instance, the stationary moments for  $y$ ,  $cy$ , and  $y'y$  can be written as

$$E[y] = \sum_{r=1}^2 \sum_{i=1}^n \rho_{ir}^* y(k_i, \xi_r)$$

$$E[cy] = \sum_{r=1}^2 \sum_{i=1}^n \rho_{ir}^* c(k_i, \xi_r) y(k_i, \xi_r)$$

$$E[y'y] = \sum_{s=1}^2 \sum_{j=1}^n \sum_{r=1}^2 \sum_{i=1}^n p_{ir, js} \times \rho_{ir}^* y'(k'_j, \xi'_s) y(k_i, \xi_r)$$

### B. Calibration Procedure and Results

The model to be used for the applied general-equilibrium analysis is patently simplistic; severe restrictions have been imposed on the forms of tastes, technology, and particularly on the stochastic structure of the economy. It will be interesting to see, therefore, how such a stylized artificial economy will be able to mimic the salient features of U.S. business cycles.

Before proceeding, the time length of a period in the model has to be defined. Given that the shocks represent technical innovation in the production of new capital goods, it seems appropriate to consider annual intervals since the frequency of such developments is very unlikely to be higher. Also, the use of annual data has the advantage of avoiding seasonality issues.

Values for the model's tastes and technology parameters were chosen in the following manner. Two of the parameters could be assigned numbers straightforwardly. Following Kydland and Prescott (1982), the discount factor,  $\beta$ , was specified to be .96.

Next, capital's share of national income had an average annual value of .29 over the 1950–85 period, so this value was picked for  $\alpha$ .

The value of  $1/\theta$  corresponds to what in the literature is called the intertemporal elasticity of substitution in labor supply. Since the present model is based on a representative household, the empirical counterpart of  $1/\theta$  should summarize the variation in labor supply of all members of such a unit, both at the intensive and extensive margins. An estimate of this type, however, is not available. For adult males, Thomas Macurdy (1981) obtained estimates of about .3. From James Heckman and Thomas Macurdy (1980, 1982) the corresponding value for females is about 2.2. The first study refers to the intensive margin only (and does not include men younger than 25 years, who are likely to have higher variability in labor supply). The much higher estimate for females reflects both margins and therefore perhaps is more appropriate for current purposes. Hence, within the .3–2.2 range 1.7 was taken as a reasonable value, implying  $\theta = .6$ . Some analysis of the sensitivity of the results to the value of this parameter was carried out.

The empirical magnitude of the coefficient of relative risk aversion,  $\gamma$ , is somewhat controversial. Therefore two alternative values,  $\gamma = 1.0$  (actually  $\gamma = 1.001$ ) and  $\gamma = 2.0$ , were used. The first is close to the values found by Lars Hansen and Kenneth Singleton, 1983, and the second in accord with the estimates of Irwin Friend and Marshall Blume (1975).

The literature does not provide any guide for assigning a magnitude to  $\omega$ —the elasticity of depreciation with respect to utilization. The value of 1.42 was chosen for this parameter because, given the above value for  $\beta$ , it implied a depreciation rate of .1 in a deterministic steady state for the model. This is the depreciation rate used by Kydland and Prescott.

The only “free” parameters are those delimiting the stochastic structure of the model. To make things more manageable, let  $\pi_{11} = \pi_{22} = \pi$ , and  $\xi_1 = -\xi_2 = \sigma$ . It is easy to check that the asymptotic standard deviation and the first-order autocorrelation coefficient as-

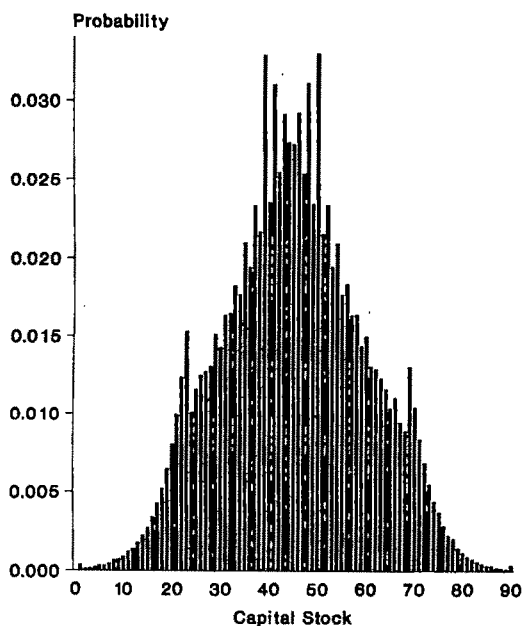


FIGURE 1. MARGINAL PROBABILITY DENSITY FOR CAPITAL

sociated with the shock,  $\xi$ , are given by  $\sigma$  and  $\lambda = 2\pi - 1$ , respectively. Thus, there are two free parameters to choose regarding the shock: its standard deviation,  $\sigma$ , and first-order autocorrelation coefficient,  $\lambda$ . As mentioned above, these parameters are to be determined so that the model generates the same standard deviation and first-order serial correlation for output as is observed in the data.

Finally, an evenly spaced grid of admissible capital stock values was chosen, which captures the ergodic set of capital stocks linked with the stochastic steady state. The grid was refined until further subdivisions did not numerically affect the covariance structure of the endogenous variables. For instance, a grid of 90 evenly spaced capital stock points spanning the interval  $[0.470, 0.667]$  turned out to be sufficient for the case where  $\gamma = 2$ ; the expected value of the capital stock in the stochastic steady state was 0.569 with a standard deviation of 0.032. The resulting unconditional probability density obtained for the capital stock is shown in Figure 1. (Note that  $\text{prob}[k = k_i] = \text{prob}[k = k_i, \xi = \xi_1] + \text{prob}[k = k_i, \xi = \xi_2]$ ).

The variables to be studied are output, consumption, investment, hours, productivity, the capital stock, and the utilization rate. Actual U.S. data, however is only available for the first five. The series used for these five variables are GNP, total consumption, total capital formation (all in 1982 dollars), average weekly hours times total employment and GNP divided by total hours. The weekly hours and employment series are household data from the *Current Population Survey*. The two remaining theoretical variables, the capital stock and the utilization rate, do not have direct empirical counterparts. The existing measure of the capital stock is constructed using the Perpetual Inventory Method which adds gross investment in constant prices each period and assumes a constant depreciation for each capital asset. In the present model, by contrast, the depreciation rate varies with utilization and new investment goods are added to the capital stock multiplied by their stochastic productivity. These differences should be important for the cyclical behavior of capital. A satisfactory counterpart to the rate of capital utilization was also not found. Existing data refer to manufacturing only, and they are calculated by comparing actual output and constructed full-capacity output indices—the latter being based on “trend-through-peak” procedures. These figures convey, therefore, similar information as a detrended manufacturing production series. For the purposes of assessing the present model these figures of utilization rates would not be appropriate since they would probably just reflect the cyclical behavior of manufacturing output.

Since by construction the variables generated by the model are stationary in levels, it is necessary to detrend the data in order to fit the model. The procedure adopted is to detrend the logged variables by a linear-quadratic time trend. Since the model is constructed for a representative agent all variables used were initially divided by the adult population. The sample period used is 1948–85.

Panel I of Table 1 portrays the statistics calculated with actual data: (1) the standard deviations, (2) the first-order serial correla-

TABLE 1—STANDARD DEVIATIONS, AUTOCORRELATIONS, AND CORRELATIONS WITH OUTPUT: U.S. DATA AND MODEL

Variables	I Annual U.S. data 1948–85			II Model $\gamma = 1$			III Model $\gamma = 2$		
	(1) <sup>a</sup>	(2) <sup>b</sup>	(3) <sup>c</sup>	(1) <sup>a</sup>	(2) <sup>b</sup>	(3) <sup>c</sup>	(1) <sup>a</sup>	(2) <sup>b</sup>	(3) <sup>c</sup>
Output	3.5	0.66	1.00	3.5	0.66	1.00	3.5	0.66	1.00
Consumption	2.2	0.72	0.74	2.3	0.95	0.50	2.2	0.94	0.79
Investment	10.5	0.25	0.68	14.7	0.44	0.85	11.6	0.50	0.90
Hours	2.1	0.39	0.81	2.2	0.66	1.00	2.2	0.66	1.00
Productivity	2.2	0.77	0.82	1.3	0.66	1.00	1.3	0.66	1.00
Capital Stock				5.9	0.98	0.56	5.6	0.99	0.52
Utilization Rate				5.9	0.48	0.56	6.0	0.52	0.61

Note. The U.S. original data was divided by the 16+ population, then logged and detrended by a linear-quadratic time trend. Output is GNP, and consumption and (gross) investment are the totals from the national income accounts, all in 1982 dollars. Hours data are from the *Current Population Survey* (which is a survey of households) and was calculated by multiplying total employment by average weekly hours.

<sup>a</sup>(1) = standard deviations, measured in percent.

<sup>b</sup>(2) = first-order autocorrelations.

<sup>c</sup>(3) = correlations with output.

tions, and (3) the correlations with output. Panels II and III show the same statistics from the model, assuming that  $\gamma = 1$  and  $\gamma = 2$ , respectively. As discussed above, the statistics from the model were obtained using the calculated limiting joint distribution of the capital stock and technology shock along with the transition matrix, from which the limiting and conditional distributions of all other variables can be calculated.

The calibration of the model is carried out by fitting the standard deviation and serial correlation of output to the corresponding values from the data using the shock parameters  $\sigma$  and  $\lambda$ . The resulting chosen values for these parameters are the following: for  $\gamma = 1$ ,  $\sigma = .0500$  and  $\lambda = .47$ ; for  $\gamma = 2$ ,  $\sigma = .0515$  and  $\lambda = .51$ .

It is interesting to note that the required exogenous persistence in the investment shock in this model is much lower than the comparable persistence of the production function shocks following from the Solow type of growth accounting, as reported by Gary Hansen (1985) and Edward Prescott (1986). Fitting their approach to the data produces exogenous shocks that are close to a random walk. Taking Hansen's (1985) estimate for quarterly autocorrelation of the

production function shocks of .95 yields a 4-quarter correlation of .81, which is still much higher than the annual autocorrelation of 0.47 and 0.51 obtained here.

The comparison of the standard deviations of the shocks is less straightforward. This is so, not only because Kydland-Prescott (1982) and Hansen (1985) use quarterly intervals whereas annual intervals are used here, but also because of the different detrending procedures. The standard deviation of actual output from trend obtained here is 3.5 percent, while it is about 1.8 percent (at annualized rates) in Hansen (1985) and Prescott (1986), who use a more flexible notion of trend that tracks actual output movements closer. However, one can compare the ratios of the required percentage standard deviation of the exogenous shocks to that of output in the two models. Here, the ratio is  $5.15/3.50 = 1.47$  for  $\gamma = 2$ . In Hansen (1985) there are two values for this ratio, namely, 1.3 and 1.7.<sup>11</sup> Hence, on this account there

<sup>11</sup>The first value corresponds to the case with indivisible labor and the second with divisible. The standard deviation of the shock  $x_t = \rho x_{t-1} + \mu_t$  is  $\sqrt{\sigma_u^2/(1 - \rho^2)}$ , where  $\sigma_u$  was equal to .00712 and .00929 for the two cases and  $\rho = .95$ ; Hansen (1985).



seems to be no advantage to the present model.

However, there is indeed an advantage when the meaning of the exogenous shock in this framework and in the Kydland-Prescott, 1982, Hansen 1985 economies are taken into account. In the latter models, the shock affects the productivity of the *entire* capital stock and labor inputs. In the present model the shock refers only to the productivity of *new* capital goods. Since productivity changes related to new capital are perhaps more plausible than overall productivity changes, a shock of a given magnitude seems a weaker requirement in this model than in the models where the shock applies to all the existing capital stock and labor.

An inspection of Table 1 will now commence. The most salient feature of the standard deviations of the actual data, shown in column (1) of panel I, is the well-known fact that investment is much more volatile than output, and consumption is less. In column (1) of panels II and III it can be seen that, in general, the model qualitatively mimics this behavior and quantitatively exaggerates it.

Column (2) of panel I describes the persistence of the movements in the different variables, as characterized by their first-order serial correlations. Consumption and productivity have the highest autocorrelations, and investment the lowest. The model also performs fairly well in this respect. In column (2) of panels II and III consumption has the highest autocorrelation, productivity the second, and investment the lowest.

The actual correlations with output appear in column (3) of panel I. Productivity and hours have the highest correlation with output but the other variables, particularly consumption, are fairly close. This feature is reproduced by the model simply because by construction there is perfect correlation of hours with output.<sup>12</sup> The procyclical behavior of consumption, however, is highly dependent on the value of  $\gamma$ . When  $\gamma = 1$ , the correlation of consumption with output is

only .50. For  $\gamma = 2$ , this correlation increases to .79, closer to the actual value of 0.74. Also, increasing  $\gamma$  from 1 to 2, which corresponds to reducing the amount of intertemporal substitution, lowers the standard deviation of investment from 14.7 to 11.6 percent, much closer to the actual data value of 10.5 percent. Overall, if this exercise is used to choose the risk-aversion parameter  $\gamma$  from the values 1 and 2, the best fit would correspond to  $\gamma = 2$ .

To check the sensitivity of the results to the labor supply elasticity parameter (chosen to be 1.7), the figures in panels II and III were computed again using alternative values. The resulting moments (not shown) are in general very similar to those shown in Table 1.<sup>13</sup>

## V. Concluding Remarks

This paper addressed the macroeconomic effects of direct shocks to investment in a framework where the utilization rate of installed capital is endogenous. The shocks considered take the form of technological changes that affect the productivity of new capital goods only.

The results in the paper suggest that a variable capacity utilization rate may be important for the understanding of business cycles. It provides a channel through which investment shocks via their impact on capacity utilization can affect labor productivity and hence equilibrium employment. Such a mechanism may allow for a smaller burden to be placed on intertemporal substitution in generating observed patterns of aggregate fluctuations.

Because of the variable capacity utilization the model predicts the Keynesian type result of less than "full-capacity equilibrium." Unlike in the Keynesian model, however, the labor market always clears and partial capacity utilization is socially opti-

<sup>12</sup>This is a consequence of the log-linear structure of production and depreciation, and the way the only stochastic shock was introduced.

<sup>13</sup>There was one expected change though. Holding constant the percentage standard deviation and first-order autocorrelation coefficient of output, the standard deviation of hours increases to 2.3 for  $1/\theta = 2$  and declines to 2.0 for  $1/\theta = 1.4$  (for both values of  $\gamma$ ).

mal. Ironically, even if the labor market is in continuous equilibrium, it is Keynes' notion of user cost that generates a Keynesian type of expansionary effect of investment shocks on employment.

#### APPENDIX

Consider a version of the model developed in the text which still incorporates shocks to the marginal efficiency of investment, but where the rate of capacity utilization is held fixed and the utility function  $\underline{U}$  is of the standard general form (i.e., not restricted to  $\underline{U}(c_t, l_t) = U(c_t - G(l_t))$ ). It will now be established that in such a framework consumption covaries *negatively* with investment and labor supply when shocked by shifts in the marginal efficiency of newly produced capital.

In the setting just described the representative agents' dynamic programming problem is given by

$$V(k_t; \varepsilon_t) = \max_{(c_t, k_{t+1}, l_t)} \left[ \underline{U}(c_t, l_t) + \beta \int_Q V(k_{t+1}; \varepsilon_{t+1}) d\Phi(\varepsilon_{t+1} | \varepsilon_t) \right]$$

subject to

$$c_t = F(k_t, l_t) - [k_{t+1} - k_t(1 - \delta)] / (1 + \varepsilon_t).$$

The efficiency condition associated with the agent's consumption/leisure decision is

$$\underline{U}_1(c_t, l_t) F_2(k_t, l_t) = -\underline{U}_2(c_t, l_t).$$

Now suppose in period  $t$  that  $\varepsilon_t$  rises. By using the above efficiency condition in conjunction with the fact that  $i_t = F(k_t, l_t) - c_t$ , two restrictions can be seen to constrain the equilibrium movements of  $c_t$ ,  $l_t$ , and  $i_t$ .

$$(A1) \quad \frac{-[\underline{U}_{11}(-\underline{U}_2/\underline{U}_1) + \underline{U}_{21}]}{\{[\underline{U}_{22} + \underline{U}_{12}(-\underline{U}_2/\underline{U}_1)] - \underline{U}_2 F_{22}/F_2\}} \times \frac{dc_t}{d\varepsilon_t} = \frac{dl_t}{d\varepsilon_t}$$

$$(A2) \quad \frac{dl_t}{d\varepsilon_t} = \frac{[\underline{U}_{11}(-\underline{U}_2/\underline{U}_1) + \underline{U}_{21}]}{\{[\underline{U}_{11}(-\underline{U}_2/\underline{U}_1)^2 + 2\underline{U}_{12}(-\underline{U}_2/\underline{U}_1) + \underline{U}_{22}] - \underline{U}_2 F_{22}/F_2\}} \times \frac{di_t}{d\varepsilon_t}.$$

Thus, a shock to the marginal efficiency of newly produced capital which causes current investment to rise will lead to an expansion of current labor effort and a *fall* in current consumption, provided that consumption and leisure are normal goods, that is,  $[\underline{U}_{22} + \underline{U}_{12}(-\underline{U}_2/\underline{U}_1)]$  and  $[\underline{U}_{11}(-\underline{U}_2/\underline{U}_1) + \underline{U}_{21}] < 0$ .

When, as in the text, the utility function is restricted to have the form  $\underline{U}(c_t, l_t) = U(c_t - G(l_t))$  income effects have no influence on labor supply. Equation (A2) still holds in this case, but  $dl_t/d\varepsilon_t = 0$  since  $[\underline{U}_{11}(-\underline{U}_2/\underline{U}_1) + \underline{U}_{21}] = 0$ . Equation (A1) no longer constitutes a restriction across movements in consumption and labor supply, but it is easy to see from the condition  $c_t + i_t = F(k_t, l_t)$  that  $dc_t/d\varepsilon_t = -di_t/d\varepsilon_t$ . Consequently, consumption still moves *negatively* with investment in response to prospective future production function shocks. A general discussion about the implications of using time-separable preferences in models of business cycles is contained in Barro and King (1984).

#### REFERENCES

- Barro, Robert J. and King, Robert G., "Time-Separable Preferences and Intertemporal Substitution Models of the Business Cycle," *Quarterly Journal of Economics*, November 1984, 99, 817-39.
- Bertsekas, Dimitri P., *Dynamic Programming and Stochastic Control*, New York: Academic Press, 1976.
- Calvo, Guillermo A., "Efficient and Optimal Utilization of Capital Services," *American Economic Review*, March 1975, 65, 181-86.
- Friend, Irwin and Blume, Marshall E., "The Demand for Risky Assets," *American Economic Review*, December 1975, 65, 900-22.
- Hansen, Gary D., "Indivisible Labor and the Business Cycle," *Journal of Monetary Economics*, November 1985, 16, 309-27.
- Hansen, Lars P. and Singleton, Kenneth J., "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns," *Journal of Political Economy*, April 1983, 91, 249-65.
- Heckman, James J. and Macurdy, Thomas, "A Life-Cycle Model of Female Labor Sup-

- ply," *Review of Economic Studies*, January 1980, 47, 47-74. [Corrigendum, October 1982, 47, 659-60.]
- Hercowitz, Zvi, "The Real Interest Rate and Aggregate Supply," *Journal of Monetary Economics*, September 1986, 18, 121-45.
- Keynes, John M., *The General Theory of Employment, Interest and Money*, London: Macmillan, 1936.
- Kydland, Finn E. and Prescott, Edward C., "Time-to-Build and Aggregate Fluctuations," *Econometrica*, November 1982, 50, 1345-70.
- Long, John B. and Plosser, Charles I., "Real Business Cycles," *Journal of Political Economy*, February 1983, 91, 39-69.
- Lucas, Robert E., "Capacity, Overtime, and Empirical Production Functions," *American Economic Review*, March 1970, 60, 23-27.
- , Prescott, Edward C. and Stokey, Nancy L., "Recursive Methods for Economic Dynamics," unpublished manuscript, Department of Economics, University of Minnesota 1985.
- Macurdy, Thomas, "An Empirical Model of Labor Supply in a Life-Cycle Setting," *Journal of Political Economy*, December 1981, 89, 1059-85.
- Merrick, John J., "The Anticipated Real Interest Rate, Capital Utilization and the Cyclical Pattern of Real Wages," *Journal of Monetary Economics*, January 1984, 13, 17-30.
- Prescott, Edward C., "Theory Ahead of Business-Cycle Measurement" *Carnegie-Rochester Conference Series on Public Policy*, Autumn 1986, 25, 11-44.
- Sargent, Thomas J., "Tobin's  $q$  and the Rate of Investment in General Equilibrium," *Carnegie-Rochester Conference Series on Public Policy*, Spring 1980, 12, 107-54.
- Taubman, Paul and Wilkinson, Maurice, "User Cost, Capital Utilization and Investment Theory," *International Economic Review*, June 1970, 11, 209-15.

# Quality, Quantity, and Spatial Variation of Price

By ANGUS DEATON\*

*In many household surveys, geographically clustered households report unit values of foods, which when corrected for quality effects and for measurement error, indicate the underlying spatial variation in prices, and can be matched to variation in demand patterns so as to estimate price elasticities. A 1979 household survey from the Côte d'Ivoire is used to estimate price elasticities for beef, meat, fish, cereals, and starches.*

In the United States, as well as in many Western European countries, income taxes are a major source of government revenue, and it is appropriate that economists should have devoted a great deal of both theoretical and empirical effort to calculating the effects of income taxes on labor supply and on government revenue. By contrast, few developing countries possess the administrative machinery that would permit a comprehensive income tax or Social Security system, so that correspondingly greater emphasis is placed upon indirect taxation and subsidies. In such circumstances, intelligent policy design requires knowledge of price elasticities for taxable commodities. Such knowledge would normally be obtained by the analysis of time-series data on aggregate demands, prices, and incomes. Unfortunately, few developing countries have time-series of a length that is adequate to estimate even own-price elasticities, let alone the cross-price elasticities that are also required. However,

many developing countries regularly collect high-quality household survey data on expenditures and quantities of a wide range of commodities. To my knowledge, such data currently exist for such diverse LDC's as Brazil, India, Sri Lanka, and Côte d'Ivoire, the Sudan, Morocco, and Indonesia, as well as for a number of developed countries, for example, the United States and the United Kingdom; a systematic search would likely reveal many more. *In principle*, these household surveys contain information on the *spatial* distribution of prices, so that if this information could be recovered in usable form, there is great potential for estimating the demand responses that are required for making policy. This paper is concerned with the development and implementation of an appropriate methodology to estimate price elasticities of demand using such cross-sectional household survey data.

In surveys where households report both expenditures and physical quantities, it is possible to divide one by the other to obtain unit values. These unit values, which depend on actual market prices, suggest that there is substantial spatial variation in prices in many developing countries, a finding that makes good sense in the presence of high transport costs. However, it is not possible to use unit values as direct substitutes for true market prices in the analysis of demand patterns. Consumers choose the *quality* of their purchases, and unit values reflect this choice. Moreover, quality choice may itself reflect the influence of prices as consumers respond to price changes by altering both quantity and quality. Measured unit values are also contaminated by errors of measurement in

\*Research Program in Development Studies, Woodrow Wilson School, Princeton University, Princeton, NJ, 08544. This paper was written while I was visiting the Welfare and Human Resources Division of the World Bank. Dwayne Benjamin provided exemplary research assistance and many helpful suggestions. Guy Laroque helped clarify conceptual issues at the beginning of the research. Helpful comments were provided by Anne Case, and by the referees of this journal. John DiNardo and Tom Lemieux pointed out an error in the calculation of the standard errors. The World Bank does not accept responsibility for the views expressed herein, which are those of the author, and should not be attributed to the World Bank or to its affiliated organizations.

expenditures and in quantities and are likely to be spuriously negatively correlated with measured quantities. In the technique developed here, market prices are treated as unobservable variables that affect quantities purchased, and that determine observed unit values with both measurement error and quality effects. Since household surveys typically collect data on *clusters* of households that live together in the same village and are surveyed at the same time, there should be no genuine variation in market prices within each cluster. Within-cluster variation in purchases and unit values can therefore be used to estimate the influence of incomes and household characteristics on quantities and qualities, and can do so without data on prices. Variation in unit values within the clusters can also tell us a good deal about the importance of measurement error. By contrast, variation in behavior *between* clusters is at least partly due to cluster-to-cluster variation in prices, and this effect can be isolated by allowing for the quality effects and measurement errors that are estimated at the first, within-cluster stage.

The plan of the paper is as follows. In Section I, I discuss previous attempts to estimate price elasticities from cross-sectional data by relating quantities purchased to their "prices," where the latter are obtained by dividing recorded expenditures by recorded quantities. The problems associated with quality choice and with measurement error are identified. I also propose a simple model of quality choice under weak separability that generates a relationship between the price and income elasticities of quality and quantity. This relationship later plays a crucial role in removing the effects of quality choice from the estimate of the price elasticity. Section II presents a model of quantity and quality choice that can be estimated and Section III contains the results for a household survey from the Côte d'Ivoire.

### I. Conceptual Background: Temptation and Its Consequences

Given the importance of estimating price elasticities in developing countries, as well as

the difficulty of doing so, it is not surprising that researchers have been prepared to make use of whatever data are available. A majority of developing countries has at some time or another conducted a household expenditure survey, usually so as to obtain weights for the calculation of a consumer price index. By definition, such surveys collect data on household expenditures, but they often go further and collect data on *quantities* purchased, particularly for foodstuffs, where quantities are well-defined, and where the consumption levels are themselves important for assessing the nutritional status of the respondents. Armed with expenditures and with quantities, the temptation to divide one by the other is irresistible. In the early classic studies by Hendrik Houthakker and Sigbert Prais, 1952, and Prais and Houthakker, 1955, the authors thoroughly analyzed the behavior of the unit values obtained by such division, but the authors were (presumably) cautious enough to resist the further temptation to use the calculated "price" to estimate price elasticities. More recently, more valorous researchers have taken up the challenge, and there have been a series of papers, by Peter Timmer and Harold Alderman, 1979; Timmer, 1981; Dov Chernichovsky and Oey Astra Meesook, 1982; and Mark Pitt, 1983, using data from Indonesia (all save Pitt) and Bangladesh (Pitt), all of which have regressed quantities on unit values, and all of which have obtained sensible and pleasing results. The dangers were perhaps more apparent than real.

Table 1 lists similar estimates for rural households from the Côte d'Ivoire for two commodities, beef and meat, the latter a broad category that includes the former. The survey and the data will be discussed in Section III below. The results given here were obtained by regressing (by ordinary least squares) the logarithm of annual quantity purchased, measured in kilos, on the logarithm of the unit value, obtained by dividing the total annual expenditure on the good by the total annual quantity purchased. In addition to the logarithm of the unit price, per capita total household expenditure on food was included as an explanatory vari-

TABLE 1—"PRICE" ELASTICITIES, BEEF AND MEAT:  
RURAL CÔTE D'IVOIRE, 1979

	Beef	Meat
No Cluster Effects	-0.56 (5.0)	-0.30 (4.3)
With Cluster Effects	-0.80 (4.6)	-0.39 (4.6)

able together with a range of household demographic indicators. There are 429 observations for beef, and 631 for meat; this is only a fraction of the total number of observations, but the (temporary) use of logarithms means that we are limited to those households that purchased positive quantities. The regression results in the first row of Table 1 are obtained by ignoring the cluster structure of the sample, while the second row involves deviations from cluster means of all variables. (Identical results would be obtained by including in the regressions dummy variables for each of the clusters.) The price elasticities are all reasonably well-determined and taken separately, either the first or second row would appear to yield satisfactory estimates. The fact that meat is less price elastic than is beef is what would be expected, given that there are substitutes for beef within the meat group as a whole. However, taken together, the results in the two rows are more disturbing. Since the model is supposedly one of spatial price variation, and since price variation within clusters should be absent, the subtraction of the cluster means should make estimation of a price elasticity impossible. In fact, the estimated price elasticity *increases*. The resolution of these (and other) problems involves a good deal of analysis and it is to this that the remainder of this section is devoted.

#### A. Quality

The unit value of "meat" is clearly not a price. Meat is not a homogeneous commodity, but a collection of commodities, in this case, agouti (a large rat), palm squirrel, venison, other game animals, game birds, chicken, guinea fowl, beef, pork, mutton, goat, and canned meat. Not all of these have the same income elasticity, so that richer households

will consume not only more than poorer households, but also in different proportions. What is generally to be expected is that the price per kilo will be higher for the goods more heavily consumed by the rich so that there will be a positive relationship between the unit value of meat and the level of household income. Even for a more narrowly defined commodity such as beef, there are more and less expensive cuts, and there are lean, scrawny (and cheap) agoutis as well as fat, sleek, and tasty ones. Houthakker and Prais (1952) give several estimates of what they call the "elasticity of quality," defined as the elasticity of unit value with respect to total household expenditure (or income); see also J. S. Cramer (1973).

One immediate consequence of this analysis is that, insofar as unit values reflect quality as well as genuine price variation, they are *chosen* by consumers just as are quantities. The regression of quantity on unit value is therefore a regression of one choice variable on another, and runs all the usual risks of possible lack of identification, simultaneity bias, and interpretational ambiguity. But there is another, and possibly more important issue. Prices will themselves affect the choice of quality. If market prices rise, consumers can not only alter the quantity that they buy, but also the quality, or more precisely the composition of their purchases within the group. If protein and calories are of greater primary concern than is "taste," then a sensible reaction to bad times is to move to less expensive cuts with little sacrifice of nutritional levels. The effect of this sort of substitution will be that an increase in price will generate a less than proportionate increase in unit value. To fix ideas, suppose that market prices for all types of meat are higher in village A than in village B, and that the per capita weight of meat consumed in A is correspondingly lower. If price were directly observed, and other variables adequately taken into account, the price elasticity could be directly estimated from the relationship between price and quantity. But if unit values are used, the same quantity difference will be ascribed to a smaller unit value difference, because of the quality effect, so that the "price elasticity" will be

exaggerated. Note that this would be true even if the econometric equations were to fit the data perfectly; the problems associated with the simultaneity of quantity and quality introduce additional complications.

Without information on all three of quantity, quality, and price, it is generally impossible to estimate everything that we want to know. Some theoretical restriction is required, and I obtain it from a simple model of the way in which quality is influenced by price. The basic assumption is that meat forms a separable branch of preferences, so that the demand for individual meats depends on the total meat budget and on the prices of the individual meats within the branch. In consequence, changes in the level of market prices of all meats together affect the demands for individual meats in exactly the same way as do changes in the total budget devoted to meat. But the "quality" of meat depends on the composition of demand within the meat group. In consequence, if we know how the quality of meat changes with changes in income, we can predict the effects of changes in absolute prices on the unit value index. As I shall show below, if the quality elasticity of meat is zero, the unit value index moves proportionately with the market price of meat. If the quality elasticity is positive, as would normally be the case, unit value will move less than proportionally with prices, the shortfall depending on the size of the quality elasticity as well as on the overall price elasticity of meat. The remainder of this subsection is devoted to a model that produces this result. The result itself will be required in the later sections to estimate price elasticities.

Suppose that at cluster (village, or site)  $c$ , the price vector for meats is  $p_c$ , where the components of  $p_c$  are the prices of the individual goods, agoutis, venison, beef, and so on. I need to assume that there exists a positive, linearly homogeneous, function of  $p_c$ ,  $\lambda_c(p_c)$ , where the value  $\lambda_c$  is to be thought of as the level of meat prices in cluster  $c$ . For example,  $\lambda_c$  could be a fixed-weight Laspeyres index, but there are many other possibilities. Given  $\lambda_c$ , I write

$$(1) \quad p_c = \lambda_c p_c^*.$$

For some purposes, it is useful to think of  $p_c^*$  as being the same for all clusters  $c$ . However, in practice the relative prices of different meats are clearly not the same in all locations; indeed not all varieties of meat will even be available in each of the clusters.

Quantities are thought of as the purchases of single individuals located at different sites, but otherwise identical. The aggregate quantity of meat purchased is  $Q_c$ , where

$$(2) \quad Q_c = k^0 \cdot q_c,$$

and  $q_c$  is the vector of meat purchases at location  $c$ . The vector  $k^0$  will be a vector of ones if it is appropriate to aggregate by weight, in which case  $Q_c$  will simply be the number of kilos of meat. I allow the more general formulation so that aggregation could be done with respect to other characteristics, for example, calories. The element-by-element ratios  $p_i^*/k_i^0$  are the prices per kilo of each of the meats, and I take these to be indicators of quality, with higher quality items costing more per kilo. Clearly, the variation in relative prices between clusters must be sufficiently limited for this interpretation to be justified.

Expenditure on the commodity group is denoted by  $E_c$ , which is  $p_c \cdot q_c$ , so that the unit value,  $V_c$ , is given by

$$(3) \quad V_c = E_c/Q_c = \lambda_c (p_c^* \cdot q_c / k^0 \cdot q_c).$$

The term in brackets, which I denote by  $v_c$ , is the measure of quality; it is the average cost per kilo at location  $c$  once price-level differences across clusters have been taken into account. It is this interpretation that limits the degree of allowable relative price dispersion across clusters. Equation (3) can be written

$$(4) \quad \ln V_c = \ln \lambda_c + \ln v_c,$$

so that, measured in logarithms, unit value is the sum of price and quality.

The next step is to use the assumption that meat is separable in preferences to derive an expression for the effect of price changes on quality. The conceptual experiment here is to vary the level of prices,  $\lambda_c$ ,

while holding fixed the within-cluster relative price structure  $p_c^*$ . By weak separability, the vector of meat demands  $q_c$  is a function of total meat expenditure and the vector of meat prices, so that

$$(5) \quad q_c = g_c(E_c, p_c) = g_c(E_c/\lambda_c, p_c^*),$$

where the second equality follows from the fact that the (group) demand functions are zero-degree homogeneous in total expenditure and prices. The quality indicator  $v_c$  is  $p_c^* \cdot q_c / k^0 \cdot q_c$ , which at fixed  $p_c^*$  and  $p^0$  is simply a function of  $q_c$ , and therefore, by (5) of the ratio  $E_c/\lambda_c$ . In consequence,

$$(6) \quad \frac{\partial \ln v_c}{\partial \ln \lambda_c} = -\frac{\partial \ln v_c}{\partial \ln E_c} + \frac{\partial \ln v_c}{\partial \ln E_c} \cdot \frac{\partial \ln E_c}{\partial \ln \lambda_c},$$

since  $E_c$ , total expenditure on meat, will itself be affected by the price change. Expenditure  $E_c$  is the product of quality  $v_c$ , price  $\lambda_c$ , and quantity  $Q_c$ , so that taking logarithms and differentiating with respect to  $\lambda_c$  gives

$$(7) \quad \frac{\partial \ln E_c}{\partial \ln \lambda_c} = 1 + \frac{\partial \ln v_c}{\partial \ln \lambda_c} + \epsilon_p,$$

where  $\epsilon_p$  is the price elasticity of the group with respect to the group price  $\lambda_c$ . Hence, from equation (6), using (7) to substitute for the last term,

$$(8) \quad \frac{\partial \ln v_c}{\partial \ln \lambda_c} = \frac{\epsilon_p \partial \ln v_c / \partial \ln E_c}{1 - \partial \ln v_c / \partial \ln E_c}.$$

Equation (8) shows that the effects of price on quality operate as income effects; an increase in the group price depresses group demand through the group price elasticity, and it is this fall in demand that generates the change in quality. The result can be expressed in the Prais-Houthakker notation by noting that, since  $v_c$  is a function of  $E_c$ ,

$$(9) \quad \frac{\partial \ln v_c}{\partial \ln x} = \frac{\partial \ln v_c}{\partial \ln E_c} \cdot \frac{\partial \ln E_c}{\partial \ln x},$$

where  $x$  is total expenditure (or income, or total food expenditure). The left-hand side is  $\eta$ , the quality elasticity as defined by Prais and Houthakker, while the last term on the right-hand side is the sum of the quality elasticity  $\eta$  and the usual quantity elasticity  $\epsilon_x$ . Making the substitutions in (9) and then in (8) gives  $\partial \ln v_c / \partial \ln \lambda_c = \eta \epsilon_p / \epsilon_x$ , so that, for unit value  $V_c$ , by equation (4),

$$(10) \quad \partial \ln V_c / \partial \ln \lambda_c = 1 + \eta \epsilon_p / \epsilon_x.$$

In consequence, if the group price changes, and we mistakenly measure the price elasticity by the relationship between the change in quantity and the change in unit value, we measure not  $\epsilon_p$  but  $\tilde{\epsilon}_p$ , where  $\tilde{\epsilon}_p = d \ln Q_c / d \ln V_c = \epsilon_p / (1 + \eta \epsilon_p / \epsilon_x)$ . This expression shows that, econometric issues apart, the comparison of quantities and unit values will tend to overstate the price elasticity in absolute magnitude, at least if, as would normally be the case, the price elasticity is negative, and the product of the price and quality elasticities is smaller in absolute value than the total expenditure elasticity. The equation also gives a means of repairing the bias, since the quality and quantity elasticities can be estimated. In Section III, I shall use the result for exactly this purpose.

### B. Spurious Correlations

The contamination of unit values by quality effects is not the only problem that lies in the way of using unit values to indicate prices, and it may not even be the most serious. Additional problems are generated by errors of measurement. Unit values are calculated by dividing expenditures by quantities, so that errors of measurement in either will not only cause the unit value to be measured with error, but will also most likely generate a spurious negative correlation between quantity and unit value. Suppose that regressions are run in logarithms, so that the point can be illustrated with a simple bivariate regression. Suppose that for observation  $i$  (either a cluster or a household), expenditure  $E_i$  and quantity  $Q_i$  are measured,



each with error. Write this

$$(11) \quad \ln E_i = \ln E_i^* + e_{1i}$$

$$\ln Q_i = \ln Q_i^* + e_{2i},$$

where true values are marked with asterisks, so that, for the unit value  $V_i$ , we have

$$(12) \quad \ln V_i = \ln V_i^* + e_{1i} - e_{2i}.$$

The errors  $e_1$  and  $e_2$  have variances  $\omega_1^2$  and  $\omega_2^2$  and covariance  $\omega_{12}$ . It is important that this covariance be taken into account; although the data recorded are expenditures and quantities, it does not follow that the measurement errors of these two magnitudes should be independent. Respondents may recall the latter by dividing the former by the price, or reconstruct expenditures from price multiplied by quantity.

Suppose that the true regression function for  $Q$  conditional on  $V$  is  $E(\ln Q_i^* | \ln V_i^*) = \alpha + \beta \ln V_i^*$ , although note that  $\beta$  is not likely to be the price elasticity, if only for the reasons in the previous subsection. If the errors are normal, the regression function of the observed quantities on observed unit values is also linear and has slope not  $\beta$ , but  $\tilde{\beta}$ , where

$$(13) \quad \tilde{\beta} = \beta(\omega_*^2/\omega_v^2) - \{(\omega_2^2 - \omega_{12})/\omega_v^2\},$$

where  $\omega_*^2$  is the true and  $\omega_v^2$  is the measured variance of  $V$ . The first term is the standard "attenuation bias" whereby  $\beta$  is multiplied by a positive factor less than unity, while the second term arises from the fact that unit values are derived by division. If unit values are correctly recalled,  $e_{1i} = e_{2i}$ , so that  $\omega_2^2 = \omega_{12}$ , and  $\tilde{\beta} = \beta$ . More generally, it seems reasonable to expect the second term to be negative, as it must be if the measurement errors in expenditures and quantities are independent. In this case, and if, as expected,  $\beta$  is negative, the two terms in (13) act in opposite directions so that the biased estimate could be either larger or smaller than the true value.

## II. Specification and Estimation

The data to be analyzed come from surveys in which the unit of observation is the household. However, such surveys are invariably geographically clustered, with clusters of a dozen or so households living in the same place and surveyed at the same time. Such a design minimizes the transport costs for the enumerators, who can spend some time in a given location, instead of constantly having to move between units that are widely separated in space. My basic assumption is that all households in the same cluster face the same market price; this price is not itself observed, but makes its presence felt in quantities purchased and in their unit values, both of which are observed. Denoting the household by  $i$  and the cluster by  $c$ , I propose two basic equations:

$$(14) \quad w_{ic} = \alpha_1 + \beta_1 \ln x_{ic} + \gamma_1 \cdot z_{ic} \\ + \theta_1 \ln p_c + f_c + u_{1ic},$$

$$(15) \quad \ln v_{ic} = \alpha_2 + \beta_2 \ln x_{ic} + \gamma_2 \cdot z_{ic} \\ + \theta_2 \ln p_c + u_{2ic},$$

where  $w_{ic}$  is the share of the budget devoted to the good (including both actual purchases and imputed expenditures),  $x_{ic}$  is the total budget,  $v_{ic}$  is the calculated unit value, and  $z_{ic}$  is a vector of household characteristics, all of which are observed. The logarithm of the cluster price,  $p_c$ , is not observed, nor is the cluster fixed-effect  $f_c$ , nor the two error terms  $u_{1ic}$  and  $u_{2ic}$ . The demand equation (14) is simply the regression function of the budget share conditional on the right-hand side variables, and should not be regarded as a structural demand equation at the level of the individual household. In particular, households that do not consume the commodity, for whom  $w_{ic} = 0$ , are included in the equation. This is the correct concept for policy analysis; if the government can impose a tax that increases all spatially dispersed prices for the good by the same amount, then the effect on revenue depends

on total demand and on the response of total demand to price, including purchasers and nonpurchasers alike. Given this, the equation is a standard Engel curve specification, linking expenditure to total outlay, price, and household characteristics. The unit value equation (15), which is observed only for households that record positive market purchases, follows the analysis of quality choice in Section I by relating unit value to the budget, to household characteristics, and to the price. The coefficient  $\theta_2$  is the response of unit value to price, and is related to the other elasticities by equation (10) above.

The share equation (14) contains a set of cluster fixed- (or random) effects  $f_c$  that represent unobservable taste variation from cluster to cluster. These are taken to be orthogonal to the unobservable price term, but they need not be orthogonal to the  $\ln x$  and  $z$  variables; they can be thought of as "residuals" in a cross-cluster explanation of purchases. Alternatively,  $f_c + u_{1ic}$  can be thought of as an error term with both cluster and idiosyncratic components. Note that the fixed effects are excluded from the unit value equation; fixed effects would preclude any inference about price from unit values, and the model would not be identified. The error term  $u_1$  has variance  $\sigma_{11}$ , and, conditional on the household making purchases in the market,  $u_2$  has variance  $\sigma_{22}$  and covariances  $\sigma_{12}$  with  $u_1$ . (Note that, because of home production, not making market purchases is not the same as having a zero-budget share.) These variances and covariances allow the model to capture the spurious relationships between quantity and price that do not come from genuine price responses; note that, in the case where the measurement errors of expenditures and unit values are independent,  $\sigma_{12}$  would be zero. More complex versions of (14) and (15) can be proposed, for example, so as to include cross-price effects, see Angus Deaton (1987), or to allow dummies to pick up broad regional taste differences that are not generated by price differences, see my working version of this paper (1986).

The parameters  $\beta_1$ ,  $\beta_2$ ,  $\gamma_1$ , and  $\gamma_2$  can be estimated by ordinary least squares applied

to the data with cluster means removed; denote these within cluster estimates by  $\tilde{\beta}_1$ ,  $\tilde{\beta}_2$ ,  $\tilde{\gamma}_1$ , and  $\tilde{\gamma}_2$ . Note that there is no selectivity problem involved in estimating the unit value equation using only those households that make market purchases; all households in the cluster face the same price, and there is no reason to link measurement error with the amount consumed. Let  $n$  be the total number of households in the  $C$  clusters, and let  $n_1$  be the number of households that record market purchases. Then if  $e_1$  and  $e_2$  are the OLS residuals from the within-cluster regressions of (14) and (15), then the variances and covariance can be estimated from  $\tilde{\sigma}_{11} = (n - k - C)^{-1} e_1' e_1$ ,  $\tilde{\sigma}_{22} = (n_1 - C - k)^{-1} e_2' e_2$ , and  $\tilde{\sigma}_{12} = (n_1 - C - k)^{-1} e_2' e_1^+$ , where  $e_1^+$  are the elements of  $e_1$  corresponding to the households that make purchases in the market. To estimate the other parameters, use the OLS estimates to "correct" the shares and unit values by calculating the two variables  $\tilde{y}_{1ic} = w_{ic} - \tilde{\beta}_1 \ln x_{ic} - \tilde{\gamma}_1 \cdot z_{1ic}$ , and  $\tilde{y}_{2ic} = \ln v_{ic} - \tilde{\beta}_2 \ln x_{ic} - \tilde{\gamma}_2 \cdot z_{1ic}$ . We are interested in the between-cluster variation in these magnitudes, so consider their cluster means,  $\bar{y}_{1,c}$  and  $\bar{y}_{2,c}$ ; from (14) and (15), the population counterparts of these magnitudes  $y_{1,c}$  and  $y_{2,c}$  satisfy

$$(16) \quad y_{1,c} = \alpha_1 + \theta_1 \ln p_c + f_c + u_{1,c},$$

$$(17) \quad y_{2,c} = \alpha_2 + \theta_2 \ln p_c + u_{2,c}.$$

It is then easy to show that, over  $C$  clusters, where cluster  $c$  has  $n_c$  households,  $n_c^+$  of which show positive consumption, we have

$$(18) \quad \text{cov}(y_{1,c}, y_{2,c}) = \theta_1 \theta_2 m_p + \sigma_{12}/n_c,$$

$$(19) \quad \text{var}(y_{2,c}) = \theta_2^2 m_p + \sigma_{22}/n_c^+,$$

where  $m_p$  is the intercluster variance of  $\ln p_c$ . Hence, if we define  $\tau$  as  $C/(\sum n_c^{-1})$ , and  $\tau^+$  as  $C/(\sum n_c^{+,-1})$ , which are the appropriate measures of average cluster size, then we take covariances over the clusters, the em-

pirical ratio

$$(20) \quad \tilde{\phi} = \frac{\text{cov}(\tilde{y}_1, \tilde{y}_2) - \tilde{\sigma}_{12}/\tau}{\text{var}(\tilde{y}_2) - \tilde{\sigma}_{22}/\tau^+}$$

will consistently estimate the ratio  $\theta_1/\theta_2$ , consistency referring to the situation where the number of clusters becomes large but the number of households per cluster remains fixed.

To understand the intuition behind this estimator, note that, if there were no "corrections" to the numerator and denominator in (20), it would be the ratio of a covariance to a variance, that is, an ordinary least squares estimator, in this case of cluster-budget share on cluster price, at least once both variables have had the effects of household characteristics netted out. There are two problems with this OLS estimator. First, there is measurement error in both share and unit value. As discussed in Section I, Part B, there is likely to be a spurious negative correlation between quantity and price, and this will result both in an inflated variance for the logarithm of the unit value, as well as in a spurious correlation between the share and the log unit value. Averaging over clusters will reduce, but not eliminate the bias induced by these effects and both the covariance and the variance have to be corrected using the error covariance and variance  $\sigma_{12}$  and  $\sigma_{11}$  estimated at the first stage. Second, since unit value responds to price with an elasticity of  $\theta_2$ , which is not in general equal to unity, budget shares respond to unit values with a coefficient  $\theta_1/\theta_2$ , rather than the coefficient  $\theta_1$  that would be obtained if  $\theta_2$  were unity. This second problem cannot be corrected from the data, but requires application of the quality model of Section I. In particular, (10) gives  $\theta_2$  as  $1 + \beta_2 \varepsilon_p / \varepsilon_x$ . In the model defined by (14) and (15), neither the price elasticity  $\varepsilon_p$  nor the expenditure elasticity  $\varepsilon_x$  are constant, but since  $w$  is the product of the unit value and the quantity divided by the budget, we have  $\partial \ln w / \partial \ln p = \theta_2 + \varepsilon_p = \theta_1/w$ , and  $\partial \ln w / \partial \ln x = \beta_2 + \varepsilon_x - 1 = \beta_1/w$ , so that, first substituting these expressions in the formula for  $\theta_2$ , and then replacing  $\theta_2$  by  $\theta_1/\phi$ ,

we get, after some rearrangement

$$(21) \quad \theta_1 = \phi \{ \beta_1 + w(1 - \beta_2) \} / (\beta_1 + w - \phi \beta_2).$$

An estimate of  $\theta_1$  can be obtained by replacing the quantities on the right-hand side by their estimates, although notice that  $\theta_1$  will generally vary with the budget share. Our estimates below will be calculated at the sample mean of the  $w$ 's. Note that, if  $\beta_2 = 0$ , so that there are no income effects on quality, then  $\theta_1 = \phi$ , and the estimator (20) requires no further correction.

In summary, there are three stages to the estimation. At the first, OLS applied to the within-cluster data yields estimates of the effects of total expenditure and household characteristics, as well as of error measurement variances and covariances. At the second state, the effects of the budget and characteristics are netted out, and cluster averages of the "corrected" budget shares and unit values calculated. A regression of shares on unit values, corrected for measurement error, yields an estimate of the ratio of the responses to price of the share and the unit value. At the third and final stage, the effect of price on the budget share is extracted from the ratio by use of the theory linking quality and quantity elasticities. It seems not to be the case that there exists any obvious instrumental variable estimator that will short-circuit the first two stages. For example, if (15) is used to express price in terms of unit value and the result substituted into (14), we obtain a relation between share and unit value in which the unit value term is correlated with the error term. The cluster dummies would seem to be likely instruments, but it is easy to see that this is equivalent to replacing individual unit values by the cluster means of unit values, and these, like the individual observations, are still contaminated by measurement error. Indeed, it is precisely this contamination that requires the use of the errors-in-variable estimator in (20) rather than OLS.

The estimator  $\tilde{\phi}$  given by (20) can be compared with the estimator  $\hat{\phi}$ , where  $\hat{\phi}$  is the same estimator with the tildes removed,

that is, the estimate with the first-stage parameters known.  $\hat{\phi}$  is a standard errors in variables estimator (see Wayne Fuller, 1987, p. 108), and its asymptotic variance is given by

$$(22) \quad v(\hat{\phi}) = (m_{22} - \sigma_{22}/\tau)^{-2} \\ \times \{ \pi^0 m_{22} + (m_{12} - \phi m_{22})^2 \} \\ (C-1)^{-1},$$

$$(23) \quad \pi^0 = m_{11} - 2m_{12}\phi + m_{22}\phi^2,$$

where  $m_{11}$  is the variance of  $y_{1,c}$  and  $m_{12}$  and  $m_{22}$  are, respectively, the covariance and variance in equation (20). This formula understates the variance of  $\hat{\phi}$  because it ignores the estimation uncertainty associated with the  $\beta$ 's and  $\sigma$ 's. The correct formula is given in the Appendix; in practice, and for these data, (22) is an extremely accurate approximation. The variance of the estimate of  $\theta_1$  calculated from (21) can be obtained by application of the "delta" method.

### III. Empirical Results

The data are taken from the *Enquête Budget Consommation*,<sup>1</sup> which collected data from a random national sample of households in the Côte d'Ivoire during the calendar year 1979. There are 1200 urban households in the sample, 522 in Abidjan, and 678 in other towns. The remaining 720 households are divided between the northern Savannah (264 households), and the coastal East and West Forest regions, with 312 and 144 households, respectively. Although the survey was a very ambitious one, not all of the data collected are now usable, and we are limited to only a fraction of them. Data are available on the value and weight of each

purchase during "last week" for 100 different foodstuffs, as well as on the volume and imputed value of "autoconsommation," foods grown, manufactured, or captured for own consumption. There are also data on a limited number of household characteristics including the ages and sexes of family members and the household location.

Each of the rural clusters contains twelve households, so that there are 26 East Forest clusters, 12 West Forest clusters, and 22 Savannah clusters. Since each of these clusters was visited during four different seasons of the year, there will generally be genuine (seasonal) price variation within each cluster. To deal with this, I treat clusters in different quarters as different clusters, so that there are effectively 240 rural "clusters" or more accurately cluster-quarters. The urban clusters are naturally much less dispersed than those in the countryside, so that urban households might easily buy commodities in clusters different from those in which they live. In consequence, I confine the presentation here to the rural results; even so, and although the results are not the same, the methodology appears to work just as well for urban households. Cluster-quarters in which no households purchase the good in the market have to be excluded since no unit value can be calculated; there are a total of 49 such clusters out of 240. Although such exclusions may generate some selectivity bias, note that the situation is much better than would be the case if we had to exclude all households that made no market purchases.

The variables in the specifications (14) and (15) are defined as follows. The income or expenditure variable is total annualized household expenditure on food divided by the number of people in the household. Expenditure on food rather than total expenditure is necessitated by absence of data on the latter, and will be theoretically acceptable if food is separable in preferences. Annualized expenditure is thirteen times the total value of purchases and imputed consumption observed over the total of the four weekly visits. The  $z$  variables are household demographics; we have data on the numbers of household members by sex in each of seven age-groups. The first thirteen  $z$  variables are

<sup>1</sup>The underlying data from this survey, as well as those from the *Living Standards Survey* used later in the paper are the property of the government of the Côte d'Ivoire. The tapes are lodged with the Welfare and Human Resources Division of the World Bank, to whom requests for access should be directed.

the ratios of each of these numbers to total household size (the effect of the fourteenth can be inferred from the intercept), and I also enter the logarithm of household size to allow for the possibility that demands are not linearly homogeneous in household numbers and total expenditure taken together. Expenditures and quantities were recorded at the purchase level, but are here aggregated so that unit values were derived by dividing total expenditures for the household in the relevant week by total quantities in kilograms for the same week. Nonpurchased quantities (home-produced or hunted goods) and the corresponding imputed expenditures were then added to market expenditures to give the total from which the budget shares are formed.

The detailed analysis here is confined to meat; summary results will be given for four other categories of expenditure, fresh fish, other fish (dried and smoked), cereals (rice, maize, millet, etc.), and starches (cassava, yams, potatoes, and plantains). Table 2 shows a preliminary analysis of the unit values that tests for regional and temporal price differences, as well as for quality effects. The vertical panels refer to regressions in which the number of included variables increases from left to right. The fourteen demographic variables as defined above are included in all of the regressions, as is  $\ln(PCX)$ , the logarithm of per capita total food expenditure. The quality elasticity of meat is small or zero, although for the urban households (not shown), there is a statistically significant but still small (0.10) effect. Greater variety is typically available in urban markets, so there is presumably more scope for a quality effect than in the countryside. The second, third, and fourth panels explore the effect of adding locational and time variables. The second pair of regressions includes dummies for the regions and for the quarters. Meat prices are very much higher in the East Forest than in either West Forest or in the omitted Savannah region. In the third panel the regions are replaced by a set of cluster dummies for the 59 rural clusters. The  $F$ -statistics in the bottom panel relate to these regressions and test for the significance of the cluster, quarter, and demographic effects.

The demographic effects are jointly insignificant, while the seasonal effects generate  $F$ -statistics that are significant at conventional levels. The cluster effects generate a very large  $F$ -value, one large enough to pass even the (very stringent) test proposed by Gideon Schwartz (1978), which in this context is that  $F$  should be larger than the logarithm of the sample size. The test itself is of no great moment, but the strength and significance of the cluster effects in the countryside is very important because it indicates the existence of the cross-cluster price variation that is necessary to make possible the estimation of the price elasticities. The fourth and final panel in Table 2 shows the quality elasticity from the within-cluster estimator; in this regression the same cluster in two different quarters is treated as two separate clusters, so that geographical and temporal dummies are fully interacted. It is this estimate that is taken forward to the calculation of the price elasticities. Estimation of the within-cluster Engel curve for meat yields an estimate of  $\beta_1$  of .052 with a standard error of .01; it is worth noting that without allowing for the cluster effects, though with regional dummies, the estimate of 0.026 is only half the size, so that the estimated expenditure elasticity from the within-cluster regression is twice as far from unity as that from the whole sample regression, 1.37 versus 1.19. Cluster effects are also important in the share equation and it is at least plausible that some of this importance comes from prices. The within-cluster share and unit value regressions have an estimated covariance,  $\sigma_{12}$ , of 0.00845, corresponding to a correlation coefficient of 0.07, so that, in this instance, the reporting errors in expenditures and unit values are close to being independent. Such a finding implies a negative covariance between the reporting errors in quantities and unit values, an implication which is confirmed if we estimate not a share log unit value pair of equations, but a log quantity/log unit value pair; see my paper (1986).

The second stage of estimation moves from within- to between-cluster analysis. Once the effects of expenditure and demographics have been removed from the shares and the unit values, the intercluster covariance between

TABLE 2—MEAN UNIT VALUES:  $\ln V$ 

Rural:	$n = 631, \ln(n) = 6.45$			
	Regression 1	Regression 2	Regression 3	"Within" Regression 4
$\ln PCX$	.088 (.037)	.030 (.038)	.056 (.045)	.065 (.042)
East Forest	—	.255 (.053)	—	—
West Forest	—	.007 (.070)	—	—
Clusters	—	—	×	×
Q2	—	-.173 (.061)	-.170 (.052)	×
Q3	—	-.142 (.059)	-.114 (.050)	×
Q4	—	-.085 (.061)	-.063 (.053)	×
$R^2$	.052	.109	.454	—

Note: Demos,  $F = 1.25$ ,  $p = .24$ ; quarters,  $F = 3.97$ ,  $p = .01$ ; clusters,  $F = 6.72$ ;  $p < 10^{-4}$ .

TABLE 3—BUDGET SHARES AND ESTIMATES OF QUALITY AND QUANTITY ELASTICITIES

	$w$	$\beta_2$	$\epsilon_x$	$\epsilon_p$
Meat	0.139	0.065 (.04)	1.305 (.09)	-0.793 (.12)
Cereals	0.201	0.040 (.02)	1.091 (.07)	-1.076 (.30)
Starches	0.310	0.023 (.04)	0.840 (.06)	-0.847 (.10)
Fresh Fish	0.050	0.026 (.02)	0.682 (.11)	-1.575 (.26)
Other Fish	0.080	0.030 (.02)	0.536 (.05)	-1.189 (.14)

shares and unit values is 0.0074, and the intercluster unit value variance is 0.3250, so that the OLS estimate of the response of the share to the log price, uncorrected for measurement error or quality effects, is the ratio of these two numbers, or 0.0227. The "average" cluster sizes  $\tau$  and  $\tau^+$  are 11.6 and 1.98, respectively, so that to correct for measurement error using (20), 0.0084 (the estimate of  $\sigma_{12}$ ) divided by 11.6 is subtracted from 0.0074, and the result divided by 0.3267 less 0.1074 (the estimate of  $\sigma_{22}$ ) divided by 1.98, giving a result of 0.0245 with an estimated standard error of 0.0173. If unit values moved one for one with price, so that  $\theta_2 = 1$ , this would be the final estimate of  $\theta_1$ , the response of the budget share to price. As it is, the estimated quality elasticity of meat is small, 0.065, so that the application of (21) makes little difference, with a final quality-corrected estimate of  $\theta_1$  of 0.0235 with a standard error of 0.0166. The corresponding estimate of  $\theta_2$  is 0.9604, so that the final price and (total food) expenditure elasticities for the quantity of meat are -0.793 and 1.305. By contrast, a direct regression of the meat share on log unit value,  $\log PCX$ , the

demographics and regional dummies yield estimates of the same elasticities of -0.450 and 0.775. The differences lie in the treatment of quality, and of measurement error, as well as in the fact that the direct regression can include only those households that record market purchases of meat and for whom a unit value index is directly available.

Table 3 shows the full set of quality and quantity elasticities for the five goods; for each, the food budget share is given first, followed by the estimates of the expenditure elasticity of quality, the expenditure elasticity of quantity, and the price elasticity of quantity. While all of these numbers seem reasonable, there is little with which to compare them, and the model developed in this paper is essentially exactly identified, so that it is difficult to explore the validity of the assumptions that lie behind the estimates. However, there now exist new, and very different data from the Côte d'Ivoire that allow some limited comparisons to be made. In 1985, the World Bank, in conjunction with the government of the Côte d'Ivoire, conducted a *Living Standards Survey* which

TABLE 4—SOME COMPARISON TOTAL EXPENDITURE AND PRICE ELASTICITIES:  
CÔTE D'IVOIRE, 1985

	$\epsilon_x$	$\epsilon_p$		$\epsilon_x$	$\epsilon_p$
Beef	1.56	-1.91	Maize	0.52	-1.19
Fish	0.74	-1.31	Yams	1.00	-1.49
Imported Rice	0.73	-1.40	Plantain	0.95	-1.41
Domestic Rice	0.73	-1.02	Cassava	0.85	-0.91

collected a large volume of household survey data. Although household expenditures were surveyed, no data were collected on physical quantities. However, a complementary survey gathered data on prices, by direct observation in the markets used by the households in the survey. It is therefore possible to associate market prices with individual households, and to examine the relationship between budget shares, household total expenditures, and the market prices. Of course, this comparison is far from perfect; without quantity data, no allowance can be made for quality effects, the market price data are also subject to considerable measurement error, and the definitions of the goods in the two surveys are not the same. However, the analysis given above suggest that the quality effects may not be very large, so the comparisons are worth making.

Table 4 shows estimates for some of the relevant goods; the numbers are elasticities with respect to total expenditure, not just food, while the price elasticities are taken holding total expenditure constant as opposed to food expenditure. In consequence, the expenditure elasticities in Table 4 ought to be rather less than those in Table 3, while if food as a whole is not very price elastic, the same will be true of the price elasticities. Allowing for the usual degree of uncertainty, the expenditure elasticities correspond rather well between the tables, while the price elasticities in Table 4 are larger than would be expected from Table 3, perhaps because most of the goods in the former are more narrowly defined. Given the difficulties of estimating price elasticities in any context, I view these results as being encouraging, although a final verdict on the method will have to await further experiments in other countries.

## APPENDIX:

## DERIVATION OF STANDARD ERRORS

This brief Appendix derives a formula for the asymptotic variance of  $\tilde{\phi}$  in equation (20) that takes into account the fact that the  $y$ 's and  $\sigma$ 's are estimated, not known. Write  $\tilde{m}_{12}$  and  $\tilde{m}_{22}$  for the covariance and variance in (20) and  $\hat{m}_{12}$  and  $\hat{m}_{22}$  for the corresponding magnitudes using the unknown  $y_{1,c}$  and  $y_{2,c}$ . Then, ignoring terms of higher order,

$$\begin{aligned} (A1) \quad & \sqrt{C}(\tilde{m}_{12} - \hat{m}_{12}) \\ &= - (C^{-1} \Sigma y_{1,c} w'_c) \{ \sqrt{C}(\tilde{b}_2 - b_2) \} \\ &\quad - (C^{-1} \Sigma y_{2,c} w'_c) \{ \sqrt{C}(\tilde{b}_1 - b_1) \}, \end{aligned}$$

where  $w_c$  are the cluster means of the variables included in the first-stage regressions, expressed as deviations around the grand mean. Using (A1) and the similar expression for  $\tilde{m}_{22}$  to expand (20) around the true value  $\phi$ ,

$$\begin{aligned} (A2) \quad & (\tilde{\phi} - \phi) = (\hat{\phi} - \phi) \\ &\quad - \{ (s_1 - 2\phi s_2)'(\tilde{b}_2 - b_2) + s_2'(\tilde{b}_1 - b_1) \} \\ &\quad - (\tau^+ m_{22}^*)^{-1} \{ (\tilde{\sigma}_{12} - \sigma_{12}) \\ &\quad \quad - \rho \phi (\tilde{\sigma}_{22} - \sigma_{22}) \}, \end{aligned}$$

where  $\rho$  is the ratio  $\tau^+/\tau$ ,  $m_{22}^*$  is  $m_{22} - \sigma_{22}/\tau^+$ , and  $s_1$  and  $s_2$  are the probability limits of  $\Sigma y_{1,c} w'_c / C m_{22}^*$  and  $\Sigma y_{2,c} w'_c / C m_{22}^*$ , respectively. The estimates of the  $b$ 's and the  $\sigma$ 's are asymptotically independent, and both are independent of  $\hat{\phi}$ , so that the variance has three terms corresponding to the three terms on the right-hand side of

(A2). Hence,

$$\begin{aligned}
 \text{(A3)} \quad v(\tilde{\phi}) &= v(\hat{\phi}) \\
 &+ \left\{ \sigma_{22}(s'_1 - 2\phi s'_2)(X'_2 X_2)^{-1} \right. \\
 &\quad \times (s_1 - 2\phi s_2) + \sigma_{11}s'_2(X'_1 X_1)^{-1}s_2 \\
 &\quad \left. - 2\sigma_{12}(s'_1 - 2\phi s'_2)(X'_1 X_1)^{-1}s_2 \right\} \\
 &+ (\tau^+ m_{22}^*)^{-2}(n - C)^{-1} \\
 &\quad \times \left\{ \pi^1 \sigma_{22} + (\sigma_{12} - \rho \phi \sigma_{22})^2 \right\},
 \end{aligned}$$

where  $(X'_1 X_1)$  and  $(X'_2 X_2)$  are the two moment matrices from the within-cluster regressions,  $\pi^1 = \sigma_{11} - 2\rho\phi\sigma_{12} + \sigma_{22}\rho^2\phi^2$ , and  $v(\hat{\phi})$  is as given by (22).

## REFERENCES

- Chernichovsky, Dov and Meesook, Oey Astra, "Patterns of Food Consumption and Nutrition in Indonesia," unpublished paper, Washington: The World Bank, 1982.
- Cramer, J. S., "Interaction of Income and Price in Consumer Demand," *International Economic Review*, June 1973, 14, 351-63.
- Deaton, Angus, "Quality, Quantity, and Spatial Variation of Price," Research Program in Development Studies Working Paper No. 127, Woodrow Wilson School, Princeton, NJ, 1986.
- , "Estimation of Own and Cross Price Elasticities from Household Survey Data," *Journal of Econometrics*, September/October 1987, 36, 7-30.
- Fuller, Wayne A., *Measurement Error Models*, New York: Wiley & Sons, 1987.
- Houthakker, Hendrik S. and Prais, Sigbert J., "Les Variations de Qualité dans les Budgets de Famille," *Économie Appliquée*, January-March 1952, 5, 65-78.
- Pitt, Mark M., "Food Preferences and Nutrition in Rural Bangladesh," *Review of Economics and Statistics*, February 1982, 65, 105-14.
- Prais, Sigbert J. and Houthakker, Hendrik S., *The Analysis of Family Budgets*, New York: Cambridge University Press, 1955.
- Schwartz, Gideon, "Estimating the Dimension of a Model," *Annals of Statistics*, March 1978, 6, 461-4.
- Timmer, C. Peter, "Is There 'Curvature' in the Slutsky Matrix?," *Review of Economic Statistics*, August 1981, 63, 395-402.
- and Alderman, Harold, "Estimating Consumption Parameters for Food Policy Analysis," *American Journal of Agricultural Economics*, December 1979, 61, 982-87.



# On the Organization of Rural Markets and the Process of Economic Development

By ALLAN DRAZEN AND ZVI ECKSTEIN\*

*How does the organization of rural land and labor markets affect capital accumulation and long-run aggregate income in the development process? We show that in a simple dual economy model capital accumulation and aggregate income will be lowest when both factor markets in agriculture are fully competitive, higher when land is not traded but the labor market is competitive, and may be highest in the absence of competitive markets in both factors in the agricultural sector.*

The dual economy growth model (Arthur Lewis, 1954; Gustav Ranis and John Fei, 1961; Dale Jorgenson, 1961; and Avinash Dixit, 1973) is thought to provide a good description and tool of analysis for problems of development. The sectoral division chosen reflects several key distinctions between the agricultural and manufacturing sectors. The main one of course has been product specialization, the agricultural sector producing food, used solely for consumption, the industrial or manufacturing sector producing goods which may be used for either consumption or investment.

Product specialization is not the only difference between the two sectors, however. Factor inputs and methods of production are quite different, as is the location of the two sectors, agriculture of course being predominantly rural, manufacturing predomi-

nantly urban. The economic and social organization of the two sectors can be quite different as well. We find a number of countries in which the manufacturing sector is mainly competitive or "capitalist," while the rural sector is largely characterized by non-competitive land and labor markets, a description common to many models of development.

A central question which development models address is the transition from a low-income rural economy to a higher-income urban or manufacturing economy. Typically, the focus of interest has been on a positive description of the dynamics of the economy or on government policies to foster capital accumulation, which is the main source of growth, taking as given the basic characteristics set out above. Specifically, the literature has emphasized the role of rural income and the agricultural surplus in affecting the migration of labor and the growth of the economy. Lewis, 1954, and Ranis and Fei, 1961, emphasized the need for surplus labor in agriculture, while Jorgenson, 1961, stressed the effects of rural income and food supply in inducing migration to the urban sector.

The focus of this paper is quite different. Rather than considering only a single type of organization of the rural sector, we ask how changes in its organization will affect the process of development. More specifically, we ask how the organization of rural factor markets will affect saving and the accumulation of capital in the short- and long run. We

\*University of Pennsylvania, Department of Economics, Philadelphia, PA, 19104, and Tel-Aviv University; and University of Pittsburgh, Department of Economics, Pittsburgh, PA, 15620, and Tel-Aviv University, respectively. We wish to thank Jon Eaton, John Harris, Elhanan Helpman, Robert Pindyck, Efraim Sadka, Neil Wallace, and seminar participants at Minnesota, Tel-Aviv, Yale, and the Institute for International Economic Studies, Stockholm. A part of this paper was written while the first author was visiting the IIES, which he wishes to thank for its warm hospitality. Financial support from the David Horowitz Institute for Economic Development and the Foerder Institute for Economic Research, Tel-Aviv University, is gratefully acknowledged.

will look at rural land and labor markets and compare the competitive case (that is, freely traded factors being paid their marginal products) with the case where markets are noncompetitive or nonexistent.

Our interest in the organization of the rural sector is motivated by, among other things, the question of land and other sorts of reforms in the rural sector. Specifically, the argument for more equal distribution of land or competitive payments to labor is that these will increase welfare of rural workers. While a given reform may clearly increase worker welfare in a static model where factor supplies to each sector are fixed, whether the same will be true in a dynamic model in the longer run will depend on how factor supplies are affected. This means considering both the process of equilibrium migration from rural to urban sector and the process of capital accumulation. If a given reform significantly affects capital accumulation, its long-run effect on welfare may be quite different from its short-run effect. The main result of this paper is to show that in a simple growth model, the steady-state capital stock may be lower when rural land and labor markets are competitive than when competitive markets for either or both of these factors are absent. This suggests that any evaluation of rural reform should be done in an explicitly dynamic model.

We consider a market-clearing, overlapping generations model with saving and capital accumulation. Migration thus becomes an equilibrium phenomenon, with workers migrating to equalize wages between the rural and the urban sector. To highlight our interest in the saving process and the land market, we will assume that there is no population growth, no technical progress,<sup>1</sup>

<sup>1</sup>In models where land is fixed and essential to production, exogenous population growth and technical progress must balance each other in steady state. Our assumption, therefore, in no way changes the basic characteristics of the steady state. Jorgensen, 1961, Dixit, 1973, and Paul Zarembka, 1970, analyzed issues of exogenous technical progress, food production, population growth, and the elasticity of food consumption in affecting the development process. Here we abstract from these issues.

and that agricultural and manufacturing goods are perfect substitutes in consumption. On the production side the two sectors differ by the assumption that capital is an input only in manufacturing (and can only be produced in the manufacturing sector), whereas land is used only in agriculture. These assumptions allow us to focus on the role of rural land and labor markets in affecting capital accumulation in the urban sector.

The organization of the paper is as follows. In Section I we present the general setup of the model. Section II presents the benchmark competitive case, while in Section III we consider a model where land is not traded and compare it to the competitive economy. In Section IV we consider the case in which there are neither competitive land nor labor markets in the rural sector. In this section we also compare results of the various models in terms of the steady state and the dynamic equilibrium path of the capital stock. In Section V we analyze the optimality of the allocations that result from the exclusion of the markets. Section VI contains our summary and conclusions.

## I. The Model

We consider a model with two sectors.<sup>2</sup> The urban sector produces commodity  $Y$  using capital  $K$  and labor  $L^y$  as inputs. Output is given by

$$(1) \quad Y = G(K, L^y).$$

$Y$  can be used for consumption or investment (that is, capital accumulation). The rural sector produces (agricultural) commodity  $X$  using only land  $A$  and labor  $L^x$  with output given by

$$(2) \quad X = F(A, L^x).$$

$X$  is used only for consumption and is not

<sup>2</sup>A model extremely close in setup to this one is that of Jonathan Eaton (1987), which analyzes international trade questions. Jean Tirole (1985) carefully analyzes the role of nonproduced assets in the Diamond model.

storable. Both  $F(\cdot)$  and  $G(\cdot)$  are continuous, twice differentiable, and linear homogeneous, with positive output requiring positive inputs of both factors. Furthermore, as an input approaches zero, its marginal product approaches infinity, given a positive value of the other input. We further assume that labor is perfectly mobile between sectors with zero cost.

The total supplies of land  $A$ , initial capital  $K_0$ , and labor  $L$  are exogenously given. Hence, the production of the agricultural good can change only with changes in labor input in that sector and is bounded above by the total supply of land and labor.

Population consists of  $L$  people in each generation, each of whom lives for only two periods. In each generation at time  $t$   $L_t^x$  people are working in the rural sector and  $L_t^y (= L - L_t^x)$  in the urban sector.  $e_t^x$  and  $e_t^y$  are the fractions of the total population in the rural and urban sectors at  $t$ . All workers are homogeneous in skills and preferences. For simplicity, we assume that  $X$  and  $Y$  are perfect substitutes in consumption. Total consumption at age  $i$  ( $i=1,2$ ) in period  $t$  for an individual is defined as

$$(3) \quad c_t^i = x_t^i + d_t^i,$$

where  $x_t^i$  and  $d_t^i$  is individual consumption of the agricultural and manufacturing goods.

Perfect substitutes imply that relative demands are perfectly elastic, or, equivalently relative prices are fixed.<sup>3</sup> Therefore, even if one sector is not competitive, production must still be on the efficient frontier.

Each person is endowed with one unit of labor in his first period of life and no labor capacity in his second period of life. The individual decision problem when young is then given by choosing total consumption in each period and savings  $s_t$  to maximize utility

$$(4) \quad U = U(c_t^1, c_{t+1}^2)$$

subject to

$$(5) \quad c_t^1 = w_t - s_t + \alpha_t^1$$

$$(6) \quad c_{t+1}^2 = R_{t+1}s_t + \alpha_{t+1}^2,$$

where  $w_t$  is wage income from work,  $\alpha_t^i$  represents possible other sources of income in the  $i$ th period of life, and  $R_{t+1}$  is one plus the interest rate in period  $t+1$ .<sup>4</sup> The first-order condition for a maximum is

$$(7) \quad \frac{U_1(\cdot, \cdot)}{U_2(\cdot, \cdot)} = R_{t+1},$$

which yields a general saving demand function for the young

$$(8) \quad s_t = s(w_t + \alpha_t^1, \alpha_{t+1}^2, R_{t+1}).$$

One can show that if consumption is normal, saving of the young is increasing in first-period income and decreasing in second-period exogenous income.

## II. The Benchmark Competitive Case

In all the economies that we analyze we assume that the manufacturing sector is competitive. The representative firm in this sector chooses  $K_t$  at time  $t-1$ , and  $L_t^y$  at  $t$  to maximize profits which are given by

$$(9) \quad \pi_t^y = q_t G(K_t, L_t^y) - w_t L_t^y \\ + (1-\delta)q_t K_t - R_t q_{t-1} K_t,$$

where  $q_t$  is the price of  $Y$  in terms of  $X$ . Since the model is deterministic, we assume perfect foresight, so that we obtain the following first-order conditions:

$$(10) \quad w_t = q_t G_L(k_t, e_t^y)$$

$$(11) \quad q_{t-1} R_t = q_t (1-\delta) + q_t G_K(k_t, e_t^y),$$

<sup>3</sup> This assumption can be interpreted as the economy being small and open to trade in the two products.

<sup>4</sup> This setup assumes a perfect consumption-loan market. Imperfections in the capital market, sometimes thought to characterize the secondary sector, are here captured in the modeling of the land market.

where  $k_t = K_t/L$  is capital per capita. Market-clearing conditions for  $Y$  imply that

$$(12) \quad k_{t+1} - (1 - \delta)k_t + d_t^1 + d_t^2 \leq G(k_t, e_t^y) = \frac{Y_t}{L}.$$

If consumption of  $Y$  is positive, our assumption on preferences implies that the price of  $Y$  will equal that of  $X$  and  $q_t$  will equal 1. If  $Y$  is not consumed, entire urban output going to capital accumulation, then  $q_t \geq 1$ , with strict inequality holding when desired  $k_{t+1}$  exceeds urban output. The price of consumption is then the price of the agricultural good. Since capital is accumulated only for future production of  $Y$  and since increased  $Y$  increases welfare only if it is consumed, zero consumption out of urban output is possible only in the short run. In the long-run steady state, consumption of  $Y$  must be positive, so that  $q$  must equal one. In early periods of development, however, the price of the urban output would be greater than that of the consumption good. For simplicity we consider economies that are sufficiently developed that some urban output is consumed, so that  $q=1$  along the path.

The economies in this paper differ with respect to the organization of the agricultural sector. As a benchmark we use the fully competitive framework, where both land and labor are fully traded. Let  $P_t$  be the price of land in terms of consumption at time  $t$ . At  $t=1$  the stock of land is divided equally among the initial population of old people. Land is purchased at time  $t$  for use in production at time  $t+1$ . The optimization problem of producers of the agricultural good  $X$  is to maximize profits in each period, namely to maximize

$$(13) \quad \pi_t^x = F(A_t, L_t^x) - w_t^x L_t^x + P_t A_t - R_t P_{t-1} A_t$$

by choice of  $A_t$  at  $t-1$  and  $L_t^x$  at  $t$ . (Writing  $P_t A_t - R_t P_{t-1} A_t$  as  $(P_t - P_{t-1})A_t - r_t P_{t-1} A_t$ , where  $r = R - 1$ , we see that the

profits from land include capital gains and are net of user cost.) The necessary conditions for a maximum are

$$(14) \quad w_t = F_L(A/L, e_t^x)$$

$$(15) \quad R_t = \frac{F_A(A/L, e_t^x) + P_t}{P_{t-1}}.$$

In the fully competitive economy both sectors face the same wage  $w_t$  and interest factor  $R_t$ .

The market-clearing condition for  $Y$  is as given above while that for  $X$  is

$$(16) \quad x_t^1 + x_t^2 = F(A/L, e_t^x) = X_t/L.$$

The other two markets that must be cleared at each date are those for labor and capital, implying

$$(17) \quad e_t^x + e_t^y = 1$$

$$(18) \quad s(w_t, R_{t+1}) = k_{t+1} + P_t A/L.$$

The equilibrium path for this economy is solved simultaneously by equations (8), (10)–(12), and (14)–(18) for given initial values of  $L$ ,  $K_0$ , and  $A$ . (Here  $\alpha^1$  and  $\alpha^2$  are identically equal to zero, since competitive factor payments exhaust total output.) This yields not only a dynamic path for  $k$  at each  $t$ , but for prices and quantities at all dates as well. An important characteristic of the dynamic equilibrium growth path of this competitive economy is that for given exogenous variables, along the path the urban labor force and the real wage are increasing as the per capita capital stock increases. (This refers to characteristics of the path, not to comparative statics.) Since the marginal product of labor must be equal in the two sectors, an increase in the capital stock induces migration to the capital-using urban sector. As land is fixed, the real wage in the rural, and hence the urban, sector rises. (Using (10), (14), and (17) and differentiating with respect to  $k$  and  $e^x$  immediately yields the result.)

Hence, the competitive equilibrium is characterized by a path consistent with the

widely accepted facts of a positive relation between growth in production on the one hand, and migration and real wages on the other. (See, for example, John Harris and Michael Todaro, 1970.) These properties of the equilibrium should be part of any model of development. Here they arise endogenously from the basic characteristics of the economy.

We now turn to the steady state of the competitive economy, which will serve as a point of comparison for the steady states of the other economies. Eaton (1987) states sufficient conditions for the existence of a steady-state allocation of this model in which both goods are produced and consumed. These conditions ensure that saving is sufficiently large so that the equilibrium path does not converge to an allocation in which land value exhausts all saving. The steady state of the competitive economy is characterized by the following five equations:

$$(20) \quad s(w, R) = k + P \cdot A/L,$$

$$(21) \quad w = G_L(k, 1 - e^x),$$

$$(22) \quad R = (1 - \delta) + G_K(k, 1 - e^x),$$

$$(23) \quad w = F_L(A/L, e^x),$$

$$(24) \quad R = \frac{F_A(A/L, e^x) + P}{P}.$$

These five equations may be solved for the five steady-state values of the endogenous variables  $k$ ,  $P$ ,  $w$ ,  $R$ , and  $e^x$ .

From (21) and (23) we can find the steady-state relation

$$(25) \quad e^x = e^x(k),$$

which has a negative first derivative (see (19a)). Substituting (25) into (24) and (23), we obtain  $w$  and  $R$  as functions of  $k$ .  $k$  is solved from equation (20). We refer to the allocation in the fully competitive economy as allocation "CE."

### III. Competitive Labor Markets with a Group of Landlords

We now consider an economy in which the rural labor market is competitive as in Section II (as of course is the urban labor market), but where land is not traded. We assume there is a subgroup of workers of unchanging size  $L^T$  who are also the owners of land. For reasons exogenous to the model, they do not sell the land but pass it on to their descendants. With constant population, we take the number of landlords to be fixed over time at  $L^T$  comprising a fraction  $e^T (= L^T/L)$  of the population. Suppose the rent from land accrues to landlords in their second period of life. Let  $\alpha_t^2$  be the income from this land so that

$$(26) \quad \alpha_t^2 = (F(A/L, e_t^x) - w_t e_t^x)/e^T.$$

Landlords choose  $e^x$  to maximize  $\alpha_t^2$ , implying that condition (14) holds as in the CE economy. We assume that landlords also work. However, condition (15) does not hold and  $P_t$  is not defined since no land market exists. Aggregate saving is now defined by

$$(27) \quad (1 - e^T)s(w_t, R_{t+1}) + e^T s(w_t, R_{t+1}, \alpha_{t+1}^2) = k_{t+1}.$$

The equilibrium path of this economy is determined by equations (8), (10)–(12), (14), (16), (17), (26), and (27). Obviously,  $k_t > 0$  for all  $t$  since capital is the only form of saving when land is not traded. As before, the urban labor force and the real wage are increasing along the path as the per capita capital stock increases.

The steady state of this economy is described by the following equations:

$$(28) \quad (1 - e^T)s(w, R) + e^T s(w, R, \alpha^2) = k,$$

$$(29) \quad w = G_L(k, 1 - e^x),$$

$$(30) \quad R = (1 - \delta) + G_K(k, 1 - e^x),$$

$$(31) \quad w = F_L(A/L, e^x),$$

where  $\alpha^2$  is defined by (26) with no time

subscript in steady state. We refer to this allocation as "AC" (*almost competitive*).

We may now compare the dynamic paths and the steady-state allocations in the two economies CE and AC. These may be summarized as:

**PROPOSITION 1:** *Consider the AC and the CE economies starting with the same initial level of capital. If along the equilibrium path in the CE economy the price of land is increasing, constant, or only slightly decreasing over time, then the value of capital per capita and the urban labor force will be higher in the AC than in the CE economy at each date. An increase in the fraction of landlords will decrease the level of capital in the AC economy, moving it closer to that of the CE economy.*

**PROOF:**

See the Appendix.

As a corollary, one immediately notes that since the price of land is constant in a steady state, the steady-state values of capital per capita and the urban labor force are higher in the AC than in the CE economy.

The intuition of this result is that the competitive economy has less capital as there exists land as a second traded asset which lowers saving available for capital accumulation. Increasing the number of landlords widens land ownership, thus lowering saving available for capital. This result is interesting for it says that the nonexistence of the land market will induce higher capital accumulation (as well as a larger urban sector). Therefore the absence of a competitive market will yield a higher level of income, though one which is unequally distributed between the two classes of owners and non-owners of land. In Section IV we further investigate this question by considering the case of reorganizing the rural labor market by redistributing rents from land among workers in the agricultural sector.

#### IV. Absence of Competitive Land and Labor Markets

We now consider an economy where neither competitive land nor labor markets

exist in the rural sector. We retain the assumption of the previous section about land distribution and the absence of a market and add to it the assumption that workers in the agricultural sector do not receive their marginal product, but rather a share  $1 - \mu$  of average product per worker (where  $\mu$  is between 0 and 1). When  $\mu = 0$  we have the case where land is divided among rural workers; a sort of total agrarian reform.  $\mu$  could be viewed as resulting from a tenancy relation in agriculture which is common in developing nations. We take  $\mu$  to be determined exogenously.

As before, migration ensures the equality of the wage between the two sectors, implying

$$(32) \quad w_t = G_L(k, 1 - e_t^x),$$

$$(33) \quad w_t = (1 - \mu) \frac{F(A/L, e_t^x)}{e_t^x}.$$

The rest of income from agriculture is divided among the  $L^T$  landlords in the economy. We assume, as before, that landlords receive this income in the second period of their lives. Though the timing of the payment of rents may appear quite innocuous, it will in fact be crucial and therefore deserves comment. In a life-cycle model saving arises from the desire to transfer income from early periods of life in which the individual receives income to later periods when he does not. The effect of rental income on individual saving and hence on aggregate capital accumulation therefore depends on whether it induces or replaces saving. To the extent that rental income in this hereditary ownership model would probably be concentrated in later periods of life, we stress the role of rents as replacing other forms of saving and assume they are received in the second period of life. This implies that

$$(34) \quad \alpha_t^2 = \mu \frac{F(A/L, e_t^x)}{e_t^T}.$$

Before characterizing the steady state, we demonstrate that the conditions for the urban labor force to grow along the dynamic path

as the capital-labor ratio grows along the path are the same as before. Equating (32) and (33) and differentiating, we obtain

$$(35) \quad \frac{de_t^x}{dk_t} = \frac{e_t^x G_{KL}}{(1-\mu)(F_L - F/e^x) + e_t^x G_{LL}}.$$

From the concavity of  $F(\cdot)$  we know that  $F_L < F/e^x$ . Hence  $de_t^x/dk_t$  is negative as long as  $G_{KL} > 0$ , and along the equilibrium path, workers migrate to the urban sector as the capital stock grows. This result is independent of the way in which the rent from land is divided between rural workers and landlords. The lower the share of rents going to workers (the larger is  $\mu$ ), the more migration there will be. This accords with common sense.

The steady-state allocation in this share economy (which we denote "SE") is characterized by

$$(36) \quad (1 - e^T)s(w, R) + e^T s(w, R, \alpha^2) = k$$

$$(37) \quad w = G_L(k, 1 - e^x)$$

$$(38) \quad R = 1 - \delta + G_K(k, 1 - e^x)$$

$$(39) \quad w = (1 - \mu) \frac{F(A/L, e^x)}{e^x}$$

$$(40) \quad \alpha^2 = \mu \frac{F(A/L, e^x)}{e^T}.$$

We may characterize the SE allocation relative to the AC (and ultimately the CE allocation) in the following propositions.

**PROPOSITION 2:** *The steady-state allocation in the SE economy is equivalent to that in the AC economy if  $1 - \mu$  is set equal to the steady-state share of labor in the agricultural sector in the AC economy, that is,  $1 - \mu^*$ . If the share of labor  $1 - \mu$  in the SE economy is greater than (less than) the competitive share, then the steady-state capital stock in the SE economy will be greater than (less than) that in the AC economy.*

**PROOF:**

See the Appendix.

Combining this result with Proposition 1, we see that the steady-state capital stock will be highest in the absence of competitive factor markets when income distribution favors rural workers over landlords ( $\mu < \mu^*$ ), next highest in the "almost" competitive economy where land is not traded, and lowest in the fully competitive economy. One may note as a special case that when all land is divided among agricultural workers ( $\mu = 0$ ), the steady-state capital stock will be higher than in the competitive and almost competitive economies. Out of steady state  $\mu_t^*$  is changing. Hence, we can write a proposition only for constant  $\mu$  which is less than  $\mu^*$ , for all  $t \geq 1$ .

**PROPOSITION 3:** *Suppose the distribution of land rents is such that  $\mu < \mu_t^*$  for all  $t > 0$ . Then, if the AC and SE economies start with the same capital stock, the capital stock in the SE economy will be greater than the capital stock in the AC economy for all future dates.*

**PROOF:**

See the Appendix.

The intuition behind this result is easy to see. Agricultural workers receive the average product of labor in the SE economy, but the marginal product of labor in the AC economy, which is lower for  $\mu < \mu^*$ . Hence agricultural wages are higher in the SE economy for  $\mu < \mu^*$ . Furthermore, the rents from land,  $\alpha^2$ , for landlords are smaller. Hence, savings of workers and landlords in the SE economy are unambiguously larger than in the AC economy.

As a corollary, one notes this is of course true for  $\mu = 0$ . Note that the case of  $\mu = 0$  is equivalent to the standard dual economy models of the type described by Jorgenson (1961) and Dixit (1973). In these models all income from agriculture is divided among the rural population. Hence, among the economies that are described here the standard dual economy, in which there is no land market, has the highest steady-state capital stock and an equal income distribu-

tion. The competitive economy in which a land market exists also has an equal income distribution but the lowest capital stock in steady state.

If one interprets land reform as a shift in income distribution toward agricultural workers and away from landowners, then we see that land reform will increase capital accumulation and income in the long run. In fact, the same result will hold in the short run, if we think of any economy along its growth path suddenly "decreeing" a decrease in  $\mu$ . This result does not accord with the standard view of development (see, for example, Simon Kuznets, 1966), which associates higher capital accumulation and growth with a more unequal income distribution, and hence sees land reforms as introducing a fairer income distribution in the short run at the expense of higher long-run growth. This analysis presents a model where these two goals need not be traded off.

The reasons for the difference in results are easy to explain. The reasoning that usually lies behind the standard result is that saving is specified in a somewhat *ad hoc* manner, with the propensity to save being zero for low levels of income and then rising as income rises. Under such a specification, a more unequal distribution of a given level of income will increase the aggregate saving rate. In this model saving was derived from a basic life-cycle model, so that the receipt of rental income in later periods of life would tend to discourage saving and hence capital accumulation. Shifting the distribution of income away from rents and toward wages received in earlier periods of life would therefore increase saving and capital accumulation.

One can now also see why the competitive economy CE has less capital accumulation than the almost competitive economy AC. When land is traded, there are two assets with which to save, so that the amount of saving going to capital accumulation is less than if land is not traded. On the other hand an increase in the group of landlords in the AC economy would decrease the level of capital accumulation. Hence, the way that the rents from land are distributed in the economy is crucial in its effects.

## V. Optimality

We now consider the optimality of the allocations of the various economies in the short and long run. The steady-state competitive allocation satisfies the condition that  $R = 1 + F_A/P > 1$ , from equation (24). Hence, the standard Koopmans-Phelps dynamic efficiency criterion implies that the CE economy has a dynamic optimal allocation of resources over time (see Bennett McCallum, 1986, for the case of land.) However, the steady state of the CE economy is not the Golden Rule. (If *nonproductive* land with a positive price were added to the Diamond model as a second asset, the CE allocation would be the Golden Rule allocation.) In order to formalize this result, we begin by considering the allocation that maximizes steady-state welfare with equal distribution across all individuals. This is given by the solution to the following maximization problem

$$(41) \quad \text{Max}_{c^1, c^2, k, e^x} U(c^1, c^2),$$

subject to

$$(42) \quad G(k, 1 - e^x) + F(A/L, e^x) - \delta k = c^1 + c^2.$$

The first-order conditions are

$$(43) \quad G_L(k, 1 - e^x) = F_L(A/L, e^x),$$

$$(44) \quad G_K(k, 1 - e^x) = \delta$$

$$(45) \quad \frac{U_1(c^1, c^2)}{U_2(c^1, c^2)} = 1.$$

Equations (43) and (44) are the conditions for maximum aggregate consumption  $c^1 + c^2$ . Equation (43) allocates labor efficiently between the two sectors. Equation (44) is the Golden Rule for this economy since population growth is zero. Equation (45) guarantees that the distribution of consumption over the life cycle is consistent with zero population growth. We denote this allocation by "GR" (Golden Rule). We first show the



relation between the GR allocation and the competitive steady-state allocation.

**PROPOSITION 4:** *The steady-state competitive allocation CE is not the Golden Rule and the per capita steady-state capital stock in the CE economy is smaller. This also implies a smaller urban labor force in the CE than in the GR allocation.*

**PROOF:**

See the Appendix.

Proposition 4 implies that the steady-state competitive allocation does not maximize utility of the representative agent, the capital stock being below the Golden Rule level. One obvious way of intervening in the land market to reach maximum steady-state utility is to tax away the physical marginal product of land and then distribute the proceeds by lump-sum transfers.<sup>5</sup> Then, the only reason for holding land would be for capital gains. Land would be traded in steady state at a zero interest rate. (Land prices could be zero, with no land traded and steady-state  $R$  greater than one.)

Proposition 1 says that in steady state  $k^{CE} < k^{AC}$ , while Proposition 4 says that  $k^{CE} < k^{GR}$ . Hence, it is possible that the steady-state allocation in the AC economy is the Golden Rule allocation. This is the case if  $R$  in equation (30) is equal to 1. In general, we know that dynamic efficiency implies that the AC economy achieves an optimal dynamic allocation only if  $R \geq 1$ . This suggests that an economy without a land market may reach the GR allocation and yield higher steady-state welfare (on average across individuals) than one with land being freely traded. (Since the existence of landlords implies *intragenerational* heterogene-

ity, it is possible that some agents may be worse off. However, since total output is higher, nondistortionary intragenerational transfers could be used to make all individuals better off at the steady state relative to the competitive case.)

In addition, for the case where the rural labor market is distorted as well, a particular set of lump-sum taxes on landlords' income from land ( $\alpha^2$ ) combined with a transfer to first-period consumption will guarantee that  $R = 1$ , implying the Golden Rule Allocation. This may be seen manipulating equations (5)–(7) for the landlords. However, this policy yields the Golden Rule only as far as production is concerned. On the consumption side, there are two groups that only get higher welfare in steady state on "average." If  $e^T = 1$ , land is divided equally among all the population, then the steady state of the AC economy with  $R = 1$  has exactly the GR allocation. Note, however, that a higher  $e^T$  implies a lower  $k^{AC}$  so that the policy guaranteeing that  $R = 1$  for a lower value of  $e^T$  will not yield the production GR under equal distribution.

The allocation of the SE economy is not optimal since the wage rate in the agricultural sector is not equal to the agricultural workers' marginal product, unless  $\mu$  is equal to  $F_L e_i^x / F(\cdot)$  at each point of time. (This last condition is, of course, impossible for  $\mu$  fixed.) In particular, if  $\mu = 0$  the wage is higher than the marginal product of labor and the economy is overaccumulating capital (Proposition 4). Hence, we find that the standard dual economy model distributes income equally among workers and generates more growth than the other economies but has an inefficient allocation.

## VI. Summary and Conclusions

The main result of this paper is that competitive land and labor markets in the agricultural sector may induce less saving in physical capital, and hence reduce the long-run income of the economy,<sup>6</sup> relative to the

<sup>5</sup> Martin Feldstein's (1977) analysis also implies that a land tax could be used to improve the allocation in the CE economy. Neil Wallace has stressed to us that the sorts of changes in organization of markets that we discuss could be mimicked by the appropriate sort of tax. For example, a 100 percent tax on land rents in the CE economy appropriately redistributed would mimic the AC economy.

<sup>6</sup> The general point is that making a market noncompetitive (for example, monopolizing supply of a factor)

case of noncompetitive or nonexistent markets. This indicates not simply that the organization of markets in the economy may have a significant effect on the economy's development in the short and long run, but that a simple move toward more competitive rural markets need not imply an increase in welfare.

Why does the absence of competitive markets "favor" capital accumulation in this model? With a land market, as was indicated above, the possibility of saving in the form of land "crowds out" capital, in exactly the way that internally held government debt in the Diamond model reduces capital accumulation and may reduce welfare even though it expands the individual's choice set. The existence of some other asset such as money would have similar implications.

Noncompetitive rural labor markets may favor capital accumulation if the move away from competitive labor markets increases labor's share in the agricultural sector and if this increase in labor's share increases saving. We contrast this to the conventional wisdom that saving will be higher with a noncompetitive rural labor market only if labor's share is relatively low with noncompetitive market organization. If landlords receive income from the ownership of land in later periods of life, a distribution of rural income favoring labor would raise saving in this sector rather than lower it. In short, the move toward competitive rural markets might both lower total saving and lower the fraction of a given volume of saving going to capital.

Of course, there are other arguments which would yield a welfare-enhancing role for a more competitive organization of markets. This paper simply makes clear that in terms of its effects on capital accumulation in a simple model, competition need not increase welfare, implying that analyzing the effects of a change in market organization must be done in the context of a fully specified dynamic model.

may be welfare improving in the long run in a dynamic model.

#### APPENDIX:

#### PROOFS OF PROPOSITIONS 1, 2, 3, AND 4<sup>7</sup>

##### PROOF OF PROPOSITION 1:

The equilibrium conditions for the two economies may be written from (20) for the competitive economy

$$(CE) \quad s(w_t, R_{t+1}) = k_{t+1} + P_t A/L$$

and from (28) for the almost competitive economy

$$(AC) \quad s(w_t, R_{t+1}) = k_{t+1} + e^T(s(w_t, R_{t+1}) - s(w_t, R_{t+1}, \alpha_t^2)).$$

To prove the proposition, we consider the position of the two curves in  $k_t - k_{t+1}$  space. For (CE) we note that for  $P_t \geq P_{t-1}$  we may write

$$\begin{aligned} P_t A/L (R_{t+1} - 1) &= \frac{A}{L} (R_{t+1} P_t - P_t) \\ &\geq \frac{A}{L} (R_{t+1} P_{t-1} - P_t) \\ &= F_A A/L \quad (\text{from (15)}) \\ &= F(A/L, e_t^X) - F_L e_t^X, \end{aligned}$$

from the linear homogeneity of  $F(\cdot)$ . This last expression equals  $e^T \alpha_t^2$  in the AC economy from (26). We may therefore write, when  $P_t \geq P_{t-1}$ ,

$$(A1) \quad \frac{P_t A}{L} \geq e^T \frac{\alpha_t^2}{R_{t+1} - 1}.$$

To evaluate this, saving in the AC economy may be written  $s(w_t, R_{t+1}, \alpha_t^2) = (c_{t+1}^2 - \alpha_{t+1}^2)/R_{t+1}$  from (6). If  $c^2$  is everywhere normal an increase in  $\alpha^2$  implies that  $c^2$  rises, so that  $s$  falls by less than  $\alpha^2/R$  rises, meaning that the sum of  $s + (\alpha^2/R)$  rises. Noting that  $s(w, R)$  is simply  $s(w, R, \alpha^2 =$

<sup>7</sup>Efraim Sadka suggested using stability conditions to prove these propositions for the general case.

0), this implies

$$s(w_t, R_{t+1}) \leq \frac{\alpha_t^2}{R_{t+1}} + s(w_t, R_{t+1}, \alpha_t^2)$$

or

$$\alpha_t^2 / (R_{t+1}) \geq s(w_t, R_{t+1}) - s(w_t, R_{t+1}, \alpha_t^2),$$

which immediately implies

$$e^T \frac{\alpha_t^2}{R-1} > e^T (s(w_t, R_{t+1}) - s(w_t, R_{t+1}, \alpha_t^2)).$$

Combining this with (A1) implies that

$$(A2) \quad P_t A/L > e^T (s(w_t, R_{t+1}) - s(w_t, R_{t+1}, \alpha_t^2))$$

for  $P_t \geq P_{t-1}$ . The strict inequality in (A2) implies that the same result will hold for  $P_t$  slightly less than  $P_{t-1}$ . Combining (A2) with (CE) and (AC) we see that, given the condition on  $P_t$ , the curve represented by (CE) as a function of  $k_t$  lies everywhere below the curve represented by (AC), as in Figure 1.

To complete the argument, stability conditions require that an increase in  $k$  increase supply of capital more than saving available for capital accumulation. Differentiating (CE) and (AC) indicates that this implies that the curves must cut the 45-degree line from above, as in Figure 1. This means that starting at the same  $k_t$ ,  $k_{t+j}^{AC} > k_{t+j}^{CE}$  for all  $j \geq 1$ .

To prove the final part of the proposition we ask what happens to the curve for AC at each  $k_t$  as  $e^T$  rises. From (26) an increase in  $e^T$  for given  $k_t$  causes  $\alpha^2$  to fall by the same percentage (expressed as a percent of its new value). Normality of  $c^2$  means  $s(w, R, \alpha^2)$  rises by less so that  $s(w, R) - s(w, R, \alpha^2)$  falls by less than  $e^T$  rises. The product  $e^T (s(w, R) - s(w, R, \alpha^2))$  therefore unambiguously rises, so that  $k_{t+1}$  along the equilibrium curve unambiguously falls. This completes the proof.

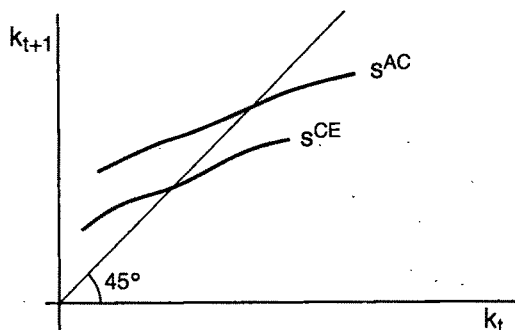


FIGURE 1.

## PROOF OF PROPOSITION 2:

The first part of the proposition is straightforward. Comparing the steady-state equations for the two economies, we see that if  $1 - \mu$  in (39) is set equal to the steady-state value of  $e^x F_L / F$  in the AC economy, then the equations and hence the allocations will be identical. Let us call this value  $\mu^*$ .

To prove the second part, consider  $\mu < \mu^*$ . At the same  $k$ ,  $R^{AC} = R^{SE}$  while  $w^{SE} > w^{AC}$ . This second relation may be demonstrated by noting that

$$w^{AC} = G_L(k, 1 - e^x) = F_L(\cdot),$$

$$w^{SE} = G_L - \frac{(1 - \mu)F}{e^x}.$$

When the production function is such that  $\mu < \mu^*$  for all values of  $e^x$ , we have  $(1 - \mu)F/e^x > F_L$  for all  $e^x$  and hence  $k$ , so that  $w^{SE} > w^{AC}$ . Finally,  $\alpha^2$  in each economy is the share of agricultural output going to capital. Therefore, by definition, if  $\mu < \mu^*$ , then  $\alpha^{SE} < \alpha^{CE}$  (where we suppress superscript 2).

Equilibrium in the capital market in both economies is represented by (28), which in steady state may be written

$$(1 - e^T)s(w, R) + e^T s(w, R, \alpha^2) = k,$$

where we denote the left-hand side, total saving of the young, as  $s^{AC}$  and  $s^{SE}$  in the two economies. Since  $w^{SE} > w^{AC}$  but  $R^{SE} = R^{AC}$ ,  $s(w^{SE}, R^{SE}) > s(w^{AC}, R^{AC})$ .

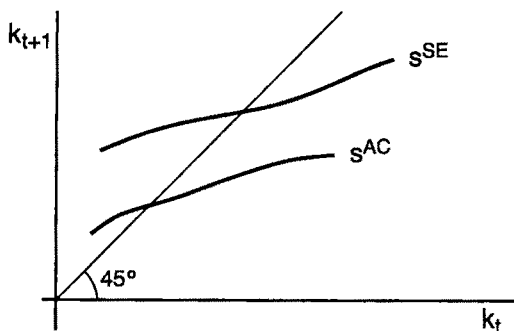


FIGURE 2

Since in addition  $\alpha^{SE} < \alpha^{AC}$ ,  $s(w^{SE}, R^{SE}, \alpha^{SE}) > s(w^{AC}, R^{AC}, \alpha^{AC})$ . For given  $e^T$ , average saving per capita,  $s^{SE}$  is therefore unambiguously greater than  $s^{AC}$  at each level of  $k$ . Stability requires that an increase in  $k$  raises steady state  $s$  less than proportionally, so that the curve cuts the  $45^\circ$  line in Figure 2 from above. Therefore,  $k^{SE}$ , steady-state capital per capita in the SE economy, is unambiguously larger than  $k^{AC}$ , steady-state capital per capita in the AC economy. This completes the proof.

### PROOF OF PROPOSITION 3:

The proof of the proposition is conceptually identical to the proof of Proposition 2. For given  $k_t$ ,  $\mu < \mu_t^*$  automatically implies that  $w_t^{SE} > w_t^{AC}$ . Similarly, at given  $k_t$ ,  $\alpha_t^{SE} < \alpha_t^{AC}$ . Therefore, for the same level of  $k_{t+1}$  (which of course is not an equilibrium),  $s^{SE}$  would exceed  $s^{AC}$ . Since  $ds/dR_{t+1} > 0$  and  $dR_{t+1}/dk_{t+1} < 0$ , the equilibrium curve (SE) must lie above (AC) in  $k_t - k_{t+1}$  space. Stability conditions require that the curve cut the  $45$ -degree line from above, as in Figure 2. Therefore, starting at the same  $k_t$ ,  $k_{t+j}^{SE} > k_{t+j}^{AC}$ , for all  $j \geq 1$ .

### PROOF OF PROPOSITION 4:

In the CE allocation  $R > 1$  due to equation (24). Hence from (7)  $c^1$  and  $c^2$  are not the same as in the GR allocation. Furthermore  $k$  and  $e^x$  cannot be the same since (22) and (24) imply that in the CE allocation  $G_K$  is greater than  $\delta$  while in the GR allocation they are equal. Equation (43) holds for both CE and GR implying that  $e^x$  is the same

function of  $k$  in both allocations, where the derivative of  $e^x$  with respect to  $k$  is negative. Hence, the function  $G_K(k, 1 - e^x)$  is the same for both allocations, and we have that

$$\begin{aligned} \frac{dG_K}{dk} &= G_{KK} - G_{KL} \cdot \frac{G_{KL}}{G_{LL} + F_{LL}} \\ &= (G_{KK}F_{LL} + G_{KK}G_{LL} - G_{KL}^2) \\ &\quad \times (G_{LL} + F_{LL})^{-1}. \end{aligned}$$

The term in the first parentheses is positive due to the strict concavity of  $G$  while the term in the second parentheses is negative.  $G_K$  is therefore decreasing in  $k$ .

### REFERENCES

- Calvo, Guillermo, "On the Indeterminacy of Interest Rates and Wages with Perfect Foresight," *Journal of Economic Theory*, December 1978, 19, 321-37.
- Diamond, Peter, "National Debt in a Neoclassical Growth Model," *American Economic Review*, December 1965, 55, 1126-50.
- Dixit, Avinash, "Models of Dual Economies," in J. A. Mirrlees and N. H. Stern, eds., *Models of Economic Growth*, New York: Wiley & Sons, 1973.
- Eaton, Jonathan, "A Dynamic Specific-Factors Model of International Trade," *Review of Economic Studies*, April 1987, 54, 325-38.
- Feldstein, Martin, "The Surprising Incidence of a Tax on Pure Rent: A New Answer to An Old Question," *Journal of Political Economy*, April 1977, 85, 349-60.
- Harris, John and Todaro, Michael, "Migration, Unemployment and Development: A Two-Sector Analysis," *American Economic Review*, March 1970, 60, 126-42.
- Jorgenson, Dale, "The Development of Dual Economy," *Economic Journal*, June 1961, 72, 309-34.
- Kuznets, Simon, *Modern Economic Growth: Rate Structure and Spread*, New Haven: Yale University Press, 1966.
- Lewis, Arthur W., "Economic Development

- with Unlimited Supplies of Labour," *The Manchester School of Economics and Social Studies*, May 1954, 22, 139-91.
- McCallum, Bennett, "The Optimal Inflation Rate in an Overlapping Generation Economy with Land," NBER Working Paper No. 1892, April 1986.
- Ranis, Gustav and Fei, John C. H., "A Theory of Economic Development," *American Economic Review*, September 1961, 51, 533-56.
- Tirole, Jean, "Asset Bubbles and Overlapping Generations," *Econometrica*, November 1985, 53, 1499-1528.
- Zarembka, Paul, "Marketable Surplus and Growth in the Low Income Economy," *Journal of Economic Theory*, June 1970, 2, 107-21.

# Public-Utility Regulators Are Only Human: A Positive Theory of Rational Constraints

By LEWIS EVANS AND STEVEN GARBER\*

*Positive public-utility models should capture incentives of regulators. Regulatory objectives are specified by appeal to standard human concerns and politics and processes peculiar to public-utility regulation. Constraints that serve the regulator are thereby derived, and connections between regulatory objectives and rules illuminated. Theoretical rationales emerge for "rate-of-return" regulation under certainty, and a largely neglected type of rate-of-return regulation under uncertainty. Motives of human regulators may explain other regulatory forms as well.*

Price and entry regulation of firms viewed as "public utilities" remains widespread, and various reforms are being considered. Assessment of changes in regulatory policies could benefit substantially by improved understanding of the actual economic effects of current policies and institutions.

Much of the positive theoretical literature on public utilities incorporates regulation simply by adding to a model of a monopoly firm one or more "regulatory constraints." A major advantage of this means of introducing the regulator is tractability. But the development of the "regulatory-constraints" literature has been plagued by controversy related to the *ad hoc* nature and consequent dubious relevance of the constraints specified. In this paper we attempt to fortify this literature by suggesting and illustrating a

general approach for specifying regulatory constraints (or rules) in a systematic fashion.

The approach is to specify an objective or "utility" function for a regulator pursuing personal ends and to *derive* constraints on the firm that serve the interests of the regulator. Constraints that well serve the (private) interests of the regulator, if they can be deduced, appear to provide a promising basis for exploiting the tractability of the regulatory-constraint approach while mitigating its greatest shortcoming. Moreover, the objectives of the regulator, being more fundamental than the constraints that a regulator would wish to place on the firm it regulates, seem to provide a more promising focus for productive debate.

In Section I we comment briefly on the theoretical literature. In Section II we present an objective function for the regulator that has as its arguments key economic outcomes of the regulatory process. In Sections III and IV we use this objective function to study the types of constraints that such a public-utility regulator would wish to impose on the firm it regulates. Section III assumes a world of certainty; uncertainty is considered in Section IV.

## I. Theoretical Literature on Public-Utility Regulation

The first, and most influential, mathematical analysis of the behavior of the regulated firm was offered by Harvey Averch and

\*Professor of Economics, Department of Economics, Victoria University of Wellington, Private Bag, Wellington, New Zealand, and Associate Professor of Economics, School of Urban and Public Affairs, Carnegie-Mellon University, Pittsburgh, PA 15213, respectively. We thank Fred Grygiel, Robert Hahn, Yehuda Koto-witz, Mark Mazur, Barry Mitnick, Edward Montgomery, Peter Navarro, Roger Noll, Thomas Romer, Glenn Woroch, participants in the Carnegie-Mellon Workshop on Political Economy and the University of Toronto Industrial Organization Workshop, and an anonymous referee for advice and comments. The usual caveat applies. Much of this research was conducted while Lewis Evans was visiting Carnegie Mellon.

Leland L. Johnson (1962).<sup>1</sup> This paper introduced an approach which has become quite common: model the regulated firm as a constrained (by regulation) monopoly. We refer to this general strategy as the "regulatory-constraint approach." In the Averch-Johnson (AJ) model the role of the regulator is represented by a constraint on the rate of return on capital. This particular regulatory constraint has been widely used<sup>2</sup> and widely criticized.<sup>3</sup> In addition, many studies have considered the AJ constraint and a firm objective other than profit maximization.<sup>4</sup> Finally, various studies have introduced uncertainty, dynamic considerations, or both.<sup>5</sup>

But the more basic issue of the usefulness of representing the role of the regulator by the device of regulatory constraint(s) has been subjected to little (if any) systematic analysis. By specifying formally the objectives of the regulator, we are able to examine this issue rigorously.

Specification of an explicit objective function for the regulator is common in the normative and the incentive-compatibility literature on public utilities.<sup>6</sup> In particular,

various theoretical studies consider maximization of some representation of "social welfare" and consider the optimal behavior (for example, pricing rule) of the regulator who pursues such an objective.<sup>7</sup> As in these literatures, we focus on deriving a type of "optimal policy" for a regulator. A fundamental difference, however, is in the nature of the objective function we specify. In pursuit of predictions concerning the economic effects of regulation involving human regulators, the objective function we employ is based on the determinants of well-being of the regulator rather than the well-being of society.

The general approach seems promising. For a particular objective function and assumptions concerning the economic environment, we are able: 1) in the case of certainty, to deduce constraints on the firm that perfectly serve the interests of the regulator, and 2) in the case of uncertainty, to deduce constraints that perfectly serve the interests of the regulator in some interesting special

<sup>1</sup>See Akira Takayama, 1969, and William J. Baumol and Alvin K. Klevorick, 1970, for clarifications and corrections, and Edward E. Zajac, 1970, and Baumol and Klevorick, 1970, for graphical interpretations.

<sup>2</sup>For example, Stanislaw H. Wellisz, 1963 (whom Alfred E. Kahn, 1970, credits with independent discovery); Elizabeth E. Bailey and John C. Malone, 1970; Zajac, 1972; Bailey, 1973; David L. McNicol, 1973; and V. Kerry Smith, 1974.

<sup>3</sup>For example, Gordon R. Corey, 1971; Paul L. Joskow, 1972a, ch. 1, 1973, 1974; Klevorick, 1973; and Joskow and Roger G. Noll, 1981, pp. 10-14.

<sup>4</sup>For example, Bailey and Malone, 1970; Zajac, 1970; Bailey, 1973; McNicol, 1973; and Smith, 1974.

<sup>5</sup>For example, Baumol and Klevorick, 1970; Bailey and Roger D. Coleman, 1971; Klevorick, 1973; E. G. Davis, 1973; Smith, 1974; Yoram C. Peles and Jerome L. Stein, 1976, 1979; Stylianos Perrakis, 1976a, b, 1983; David S. Sibley and Bailey, 1978; Nicholas Rau, 1979; H. Stuart Burness, W. David Montgomery, and James P. Quirk, 1980; Satya P. Das, 1980; Vijay S. Bawa and Sibley, 1980; and Ronald R. Braeutigam and Quirk, 1984.

<sup>6</sup>And looking beyond the public-utility literature, Sam Peltzman (1976) postulates the objective of vote maximization to study the positive economics of regulation in general. Our objective function is very different, being specified with regard to economic, political,

and procedural issues peculiar to public-utility regulation.

<sup>7</sup>Such analyses might be characterized as "normative" by those who doubt that altruism is a primary human motivation. Alternatively, they may be viewed as positive analyses for "philosopher-king" regulators. Early contributions include: Bailey, 1973, pp. 104-109; Gardner Brown, Jr., and M. Bruce Johnson, 1969; Klevorick, 1966, 1971; Hayne E. Leland, 1974; Stephen C. Littlechild, 1972; and Eytan Sheshinski, 1971. Recently the literature has focused on optimal rules for regulators in the presence of asymmetric information. (The firm generally has more information about cost and demand conditions than the regulator does.) Such studies include: David P. Baron and David Besanko, 1984; Baron and Raymond R. DeBondt, 1981; Baron and Roger B. Myerson, 1982; Jörg Finsinger and Ingo Vogelsang, 1981; Jean-Jacques Laffont and Jean Tirole, 1986; Tracy R. Lewis and David E. M. Sappington, 1987; Michael H. Riordan, 1984; Sappington, 1980, 1982, 1983; Sappington and Sibley, 1985; and Vogelsang and Finsinger, 1979. Quite recently, Joel S. Demski and Sappington, 1987, emphasizing divergence between the regulator's private interests and the interests of consumers, have considered a hierarchical model with asymmetric information: The (aggregate) consumer uses monetary incentives to motivate the regulator to gain information on the effort level of the firm, which in turn motivates the firm to exert effort that is productive in reducing costs.

cases. In addition, these results allow us to illuminate relations between regulator objectives and regulatory rules and to comment, from the perspective of the personal objectives of the regulator, on the *a priori* appeal of constraints commonly specified (in an *ad hoc* fashion) in the literature. In particular, our analysis suggests that much of the literature which abstracts from uncertainty may need only to be reinterpreted. In the case of uncertainty, however, our conclusions are much less reassuring. In particular, we conclude that almost every study (of which we are aware) that generalizes the AJ constraint to the case of uncertainty has done so in a way that seems unpromising because the constraints employed do not well serve the interests of the regulator. We propose an alternative constraint that is not subject to this criticism, and we begin the task of reconstructing.

The analysis of firm behavior subject to the constraints preferred by the regulator indicates that the tractability of the regulatory-constraints approach is preserved. In addition the analysis highlights the central role of a particular aspect of the *process* of public-utility regulation in creating inefficiency. In pursuing their interests within this process, regulators have incentives to encourage (by their choice of constraints) firms to waste resources.

## II. The Objective of the Public-Utility Regulator

The behavior of public-utility regulators seems to depend importantly on an array of economic, political, and institutional factors. Thus our objective function results from various considerations. To facilitate analysis of constraints that serve the interests of the regulator and, in turn, the economic outcomes of the imposition of such constraints, the arguments of the regulator's objective function are assumed to be economic outcomes. To specify an objective function with broad appeal, various desires of public-utility regulators are considered.<sup>8</sup> A regulator

operates in a political environment, and this suggests explicit consideration of the behavior of other participants who affect the regulator's ability to achieve his or her ends. These agents are motivated largely by economic concerns, and thus their behavior can be usefully, and most conveniently, taken to depend on economic outcomes of the regulatory process.

In order to enable consideration of issues that generally seem central to the politics of public-utility regulation, we consider two commodities with prices denoted by  $p_1$  and  $p_2$ . In this two-commodity context, the economics and political science literatures lead us to focus on four types of participants other than the regulator. These are: 1) the managers and stockholders of the regulated firm, 2) advocates for buyers of *both* of the commodities, 3) advocates for buyers of *one* of the two commodities, and 4) the courts (to which regulatory decisions can be appealed).<sup>9</sup> The objective function for the regulator is based on the assumption that each of the first three groups pursues its ends by "pressuring" the regulator by taking various actions that the regulator wishes to avoid.<sup>10</sup>

---

of unusual salience because regulators operate in a particular type of political environment (for example, prospects for reappointment or reelection, appointment or election to other offices).

<sup>9</sup>The importance of the regulated firm is amply demonstrated by the history of the "capture theory" of regulation (see, for example, George J. Stigler, 1971). Advocates of low prices in general include so-called "consumer advocates" (both independent and employed by the state), commission staff, the governor, and other politicians. See William T. Gormley (1983) for a particularly useful description of the politics of public-utility regulation and the roles played by various buyers' advocates. The importance of advocates for buyers of particular goods is also clear from Gormley's discussion. Various sources and economic impacts of asymmetric regulator concern about different prices are especially well documented by Richard A. Posner (1971), who views regulators as playing an explicitly redistributive role and provides numerous examples.

<sup>10</sup>The various groups are asymmetrically positioned to affect the various dimensions of the regulator's life, but all groups seem of general relevance. All three groups seem capable of affecting substantially the well-being of even a regulator who is largely unconcerned about some of the assumed dimensions of well-being. They seem to be able to impinge importantly, if not equally easily, on the leisure and effort levels of the

<sup>8</sup>These include standard human concerns (for example, income, leisure, effort avoidance), and also concerns



Pressure is assumed to be applied by each group in relation to its dissatisfaction with the particular outcome of the regulatory process about which each is primarily concerned.<sup>11</sup> In contrast, the role of the courts is represented by a constraint on the behavior of the regulator.

In the analysis of the certainty case, the objective of the regulator is assumed to be to maximize:<sup>12</sup>

$$(1) \quad u = u(\pi, -s, -p_1),$$

where  $\pi$  = the (economic) profit of the regulated firm, and  $s$  = the rate of return on capital of the regulated firm.

The objective function is assumed to be strictly increasing in each of its arguments, or equivalently  $\partial u / \partial \pi > 0$ ;  $\partial u / \partial s < 0$ ; and  $\partial u / \partial p_1 < 0$ .<sup>13</sup> These assumptions are based on the regulator's desire to balance optimally (various forms of) pressure exerted by the three groups described above,<sup>14</sup> and the

assumptions that these groups focus respectively on the regulatory outcomes  $\pi$ ,  $s$ , and  $p_1$ . The specification results from the following considerations.

The managers and the stockholders of the regulated firm are presumed, as is standard in economic theory, to be concerned with economic profits. Accordingly, they are assumed to apply more pressure on the regulator (i.e., do more to impinge on his or her utility) the lower is  $\pi$ . (This behavior provides the regulator with an incentive to behave as if he or she prefers higher values of  $\pi$ , other things equal.) Thus the assumption that  $\partial u / \partial \pi > 0$ .<sup>15</sup>

The specification that  $\partial u / \partial s < 0$  results from the assumption that advocates for buyers of both goods apply more pressure on the regulator the higher is  $s$ , the firm's rate of return on capital. This emphasis is based on empirical, institutional, and theoretical considerations, and is critical to our conclusions. Motivating it requires reference to aspects of the process of public-utility regulation.<sup>16</sup>

Within this process, various decisions affect importantly the general level of prices (via the calculation of "revenue requirements"). These include "rate base valuation," regulators' judgments concerning the reasonableness of operating costs, and selection of an "allowed rate of return on capital." We focus on the last of these for two com-

---

regulator through their behavior within the regulatory process. One aspect of pressure largely specific to the managers of the regulated firm has been studied empirically by Ross D. Eckert (1973, 1981): future employment opportunities with the regulated firm or industry or law firms representing industry interests, etc. This group may also be better positioned to provide financial assistance in future elections. The buyers' groups may typically be more effective in affecting future employment opportunities with "public interest" law firms, reappointment prospects, prospects for appointment to other offices, nonfinancial electoral support, and the tranquility of the regulator's existence.

<sup>11</sup>This type of behavior is postulated on the view that in an explicitly dynamic model such behavior would be rational in a broad range of circumstances. In particular, if a pressure group behaves in this fashion, regulators will (rationally) expect to benefit (other things equal) for providing an outcome more satisfactory to that group and will thus have an incentive (at the margin) to do so. The fact that production of political pressure seems widespread and costly suggests that such behavior is often rational.

<sup>12</sup>A straightforward generalization of (1) is presented in Section IV and employed to study the uncertainty case.

<sup>13</sup>The fact that  $\pi$  and  $s$  are not monotonically related in all relevant situations plays a crucial role in the analysis to follow.

<sup>14</sup>The general proposition that political agents seek to balance pressure exerted by competing sources seems

---

well grounded in various strands of both the economics and political science literatures. In a very general context, for example, Gary S. Becker (1983) emphasizes the role of political pressure applied by groups competing for economic benefits conferred through political processes. In a narrower context, in describing the Peltzman (1976) analysis, James Q. Wilson (1980, p. 361) writes that regulators seek an outcome for which the various participants are "optimally disgruntled." The perspective also seems entirely consonant with the "external signals" view of regulatory behavior described by Noll (1985). Here we consider implications of this general perspective for public-utility regulation.

<sup>15</sup>Note that the "capture theory" of regulation could be viewed as reflecting a special case of (1) with  $\partial u / \partial \pi > 0$  and  $\partial u / \partial s = \partial u / \partial p_1 = 0$ .

<sup>16</sup>See, for example, Stephen Breyer, 1982, ch. 2; Joskow, 1972a, ch. II; Kahn, 1970, ch. 2; or Charles F. Phillips, Jr., 1984, ch. 5.

plementary reasons: 1) plausibility in view of (informal) observation and theory; and 2) because doing so enables us to clarify connections between regulator objectives and regulatory rules. With regard to the former reason, the allowed rate of return is undeniably an important determinant of the general level of rates and selection of this number often involves heated public controversy (thus providing empirical support for our emphasis).<sup>17</sup> From a theoretical point of view, advocates of low prices in general may, in fact, rationally focus on the rate of return because it seems much easier to observe and influence than the regulator's judgment about the prudence of capital investments or the reasonableness of operating expenses. Influencing these other determinants of the general level of prices seems to require much more information and expertise.<sup>18</sup> With regard to connections between objectives and rules, focusing on the allowed rate of return allows us to answer the question of under what conditions regulators would find it (personally) desirable to constrain the firm's rate of return on capital (as assumed in much of the literature).

The third argument of (1) follows from the assumption that the third pressure group represents buyers of the first regulated commodity, and thus focuses on its price,  $p_1$ .

<sup>17</sup>The regulatory attention paid to the allowed rate of return has led to substantial attention by researchers, including the decision of Averch and Johnson to represent the role of the regulator entirely by a "rate of return constraint." In addition, the determination of the allowed rate of return has been the object of direct econometric study. (See, for example, Joskow, 1972b; Robert L. Hagerman and Brian T. Ratchford, 1978; R. Blaine Roberts, G. S. Maddala, and Gregory Enholm, 1978.)

<sup>18</sup>It might seem preferable to assume that advocates of low prices in general pressure directly with regard to the two prices themselves *rather than*  $s$ . Such an assumption is unattractive. Because of the process of public-utility regulation, pressuring directly on the prices and *not* on the rate of return would be largely ineffective, being vulnerable to questions such as: "Then how do you propose that we provide the firm with the required revenues previously determined within this painstaking legal process?" (Thus, even if *both* prices were specified as arguments of the regulator's objective function, it seems appropriate to include  $s$  as well.)

Consequently, we assume that this group applies more pressure the higher is  $p_1$ . The asymmetric treatment of the two prices in the regulator's objective function is introduced specifically to incorporate an apparently central element of the typical politics of public-utility regulation: The levels of some prices are much more politically sensitive than others.<sup>19</sup> Generically, we view  $p_1$  as the price of the good which is of central concern to the most politically influential interest group (or coalition of interest groups) pressuring the regulator about a particular price. It is noteworthy, however, that the central propositions of this paper would remain intact if  $p_2$  were added as a fourth argument of the regulator's utility function<sup>20</sup>.

Our reading of the literature suggests that the proposed objective function has fairly broad relevance. More fundamentally, refocusing the debate on the objectives of the regulator, and analysis on the connection between regulatory objectives and regulatory rules seem very worthwhile. It appears indisputable that more useful positive analyses of

<sup>19</sup>Many examples are available. In the case of telephones, the two goods might be local and long-distance calls. Representatives of (especially) low-income groups apply considerable political pressure to prevent prices for local calls or "basic service" from rising. Alternatively, telephone services may be decomposed into business and residential, with "consumer" advocates generally applying pressure for low residential rates. (See Noll, 1986, for a very useful discussion of the fundamental political importance of "basic service" rates in state regulation of telephone service.) In the case of electricity and natural gas, the residential vs. industrial distinction is politically very salient, pitting "consumers" against "business." Alternatively, different blocks of these commodities purchased by a particular buyer are generally priced differently. With the two commodities defined this way, the well-known controversy over declining block rates (favored by large buyers) vs. inverted block rates (favored by advocates of small, often low-income, users, and environmentalists seeking rising marginal prices to encourage energy conservation) exemplifies the importance of participants who focus on a subset of the regulated prices. Posner (1971) argues that regulated rate structures often reflect the pursuit of redistributive goals through the regulatory process, and he provides additional examples.

<sup>20</sup>See the discussion of intuition just prior to the statement of Proposition 2 in Section III.

public-utility regulation will result from explicit consideration of the objectives of the regulator, who is, after all, only human. Working with our particular objective function, we proceed to demonstrate that it is possible to derive constraints that serve the interests of the regulator, and to rationalize various forms of regulation. Moreover, we show that analysis of the behavior of the firm subject to optimal constraints is tractable. Thus our analysis holds out the hope that the tractability of the regulatory constraint approach can be preserved while working with constraints that are derived from explicit consideration of the interests of the regulator.

### III. The Certainty Case

We can now use (1) to begin to consider the circumstances under which the role of the regulator can be usefully represented by constraints on a monopoly firm, and especially constraints on the rate of return. The central premise is that regulators attempt to constrain the behavior of regulated firms in ways that benefit the regulator. We begin with the certainty case.

Assume that a profit-maximizing firm produces two goods whose prices are regulated. The demand functions are written as

$$(2a) \quad q_1^d = q_1(p_1), \text{ and}$$

$$(2b) \quad q_2^d = q_2(p_2),$$

where  $q_1^d$  and  $q_2^d$  are the quantities demanded. For convenience, we work with the firm's short-run cost function

$$(3) \quad c = c(q_1, q_2, v, k),$$

where  $q_1$  and  $q_2$  are the quantities produced,  $v$  is a vector of prices of the variable inputs, and  $k$  is the level of capital. Assuming market clearing, and letting  $r$  denote the cost per unit of capital, the firm's profits are given by

$$(4) \quad \pi(p_1, p_2, k) = p_1 q_1(p_1) + p_2 q_2(p_2) - c(q_1(p_1), q_2(p_2), v, k) - rk.$$

We assume that  $\pi(p_1, p_2, k)$  is twice continuously differentiable and strictly concave.

In order to discover regulatory constraints that serve the regulator, consider first the choices of  $p_1$ ,  $p_2$ , and  $k$  that would be made by the regulator with preferences given by (1) if that person had complete control over these choices:

$$\text{RCC:} \quad \max_{p_1, p_2, k} u(\pi, -s, -p_1)$$

$$\text{subject to} \quad \pi \geq 0,$$

where  $s \equiv (\pi + rk)/k$  is the rate of return on capital. The nonnegative-profits constraint in RCC, "which denotes regulator has complete control" is specified to represent the role of the courts, which is often associated with the *Hope Natural Gas* decision.<sup>21</sup> The function  $u(\pi, -s, -p_1)$  is assumed twice continuously differentiable, strictly quasi concave, and, as before, increasing in its three arguments.

The RCC' problem can<sup>22</sup> be rewritten as

$$\text{RCC':} \quad \max_{p_1, s} u(\pi^*(p_1, s), -s, -p_1)$$

$$\text{subject to} \quad \pi^* \geq 0,$$

where

$$(5) \quad \pi^*(p_1, s) \equiv \left\{ \max_{p_2, k} \pi \text{ subject to } sk = \pi + rk \right\},^{23}$$

and the subproblem (5) takes  $p_1$  and  $s$  as given. Various important results follow immediately from the AJ literature once RCC is written in this way.

<sup>21</sup>With its directive that allowed rates of return must "enable the company to operate successfully, to maintain its financial integrity, to attract capital, and to compensate its investors for the risks assumed..." (Joskow, 1974, p. 297).

<sup>22</sup>This decomposition of RCC merely exploits the fact  $\partial u / \partial \pi > 0$  implies that the regulator desires maximum profits for any given values of  $p_1$  and  $s$ .

<sup>23</sup>The subsequent discussions assume that the constraint in (5) is binding unless otherwise indicated.

The function  $\pi^*(p_1, s)$  is the solution to the problem studied by Averch and Johnson (1962): maximize profits over a single good subject to a constraint on the rate of return on capital. While this is a subproblem in our case, we nonetheless have:

**PROPOSITION 1:** *A regulator with objectives given by (1) and complete control over the firm's choices would want the firm to act as a profit maximizer with a constraint on the rate of return on capital.*

Proposition 1 thus provides an answer to the question of under what conditions regulators would find it personally desirable to constrain  $s$ . Moreover, letting  $\lambda$  denote the Lagrange multiplier associated with the constraint in (5), as corollaries to Proposition 1 we have for all  $s > r$ :<sup>24</sup>

**COROLLARY 1:** *The regulated firm uses more capital than the cost-minimizing level for the output it produces.*<sup>25</sup> (For example, Baumol and Klevorick, 1970, p. 166.)

**COROLLARY 2:** *If the regulatory constraint in (5) is binding:  $0 < \lambda < 1$ .* (For example, Baumol and Klevorick, 1970, p. 166.)

**COROLLARY 3:** *The derivative  $\partial k / \partial s$  is negative if  $\lambda > 0$ .* (For example, Baumol and Klevorick, 1970, p. 175.)

**COROLLARY 4:** *The derivative  $\partial \lambda / \partial s$  is negative if  $\lambda > 0$ .* (Appendix A.)

A graphical representation of RCC is presented in Figure 1, which aids subsequent discussions. The qualitative properties of the graph are established as follows. Letting subscripts denote partial derivatives, solution of RCC' requires

$$(6a) \quad u_{\pi} \pi_{p_1}^* + u_{p_1} = 0, \quad \text{and}$$

$$(6b) \quad u_{\pi} \pi_s^* + u_s = 0.$$

<sup>24</sup>See Appendix A for proofs.

<sup>25</sup>This is the famous "overcapitalization" result. In most treatments capital and labor are assumed to be the only inputs, so that the overcapitalization result is often stated in terms of the capital-labor ratio.

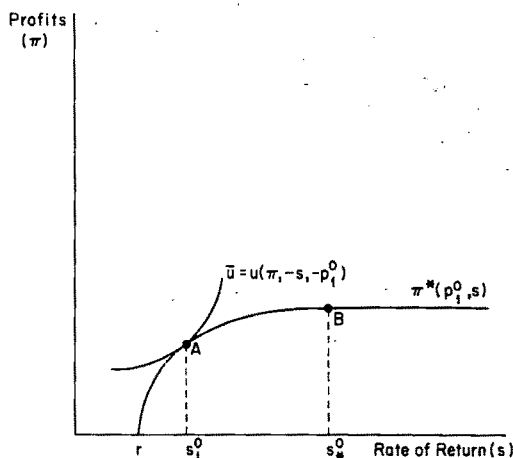


FIGURE 1. REGULATOR AND FIRM:  
A GRAPHICAL REPRESENTATION

Thus, for a given  $p_1 = p_1^0$ , say, (6b) describes the regulator's choice of  $\pi$  and  $s$ . The strict quasi concavity of  $u$  implies that the regulator's indifference curves (holding  $p_1$  fixed at  $p_1^0$ ) are shaped as in Figure 1, with utility increasing to the northwest.

The shape of  $\pi_s^*$  can be established as follows. By the envelope theorem,

$$(7) \quad \pi_s^* = \lambda k,$$

and thus  $\pi_s^* > 0$  if  $\lambda > 0$  and  $\pi_s^* = 0$  if  $\lambda = 0$ , the latter holding for all  $s$  at least as large as  $s_2^0$ , the rate of return earned by the firm if the only constraint is  $p_1 = p_1^0$ . Finally, the curvature of  $\pi_s^*$  can be verified by differentiating (7) and using corollaries 3 and 4, which establishes  $\pi_{ss}^* < 0$  for  $\lambda > 0$ , and  $\pi_{ss}^* = 0$  for  $\lambda = 0$ . If  $p_1^0$  is optimal in RCC, point A in Figure 1 represents the solution to RCC.

Intuitively, the overcapitalization results from the relationship between  $\partial \pi / \partial k$  and  $\partial s / \partial k$ , and the regulator's desire (given the level of  $p_1$ ) to adjust capital at the margin to achieve an optimal balance of the pressures due to the levels of  $\pi$  and  $s$ . In particular, at the efficient level of capital the regulator finds it personally advantageous for the firm further to increase capital, thus wasting resources. This can be seen as follows. The sign of  $\partial s / \partial k$  is that of  $(\partial \pi / \partial k - \pi / k)$ .

Assume that  $\pi > 0$  at the solution to RCC ( $\pi = 0$  is a polar case—the legal constraint in RCC is binding). Consider next that efficient use of capital involves cost minimization, which (see (4)) implies  $\partial\pi/\partial k = 0$ . Hence, at the efficient level of capital we have  $\partial s/\partial k < 0$ . Thus (unless the regulator wishes the firm to maximize  $\pi$  constrained only by the value of  $p_1$ —another polar case) the regulator desires an increase in capital from the efficient level in order to trade a lower profit (resulting in more pressure from one source) for a lower rate of return (resulting in less pressure from another source).<sup>26</sup>

Thus the regulator can achieve his or her optimum for  $u(\pi, -s, -p_1)$  by constraining the pair  $(p_1, s)$ . This provides a theoretical rationale for “rate-of-return” regulation,<sup>27</sup> and economically inefficient use of capital is implied. Consider, as an alternative, modeling the regulator as setting both  $p_1$  and  $p_2$ , as suggested by the observation that “regulators set prices not rates of return.”<sup>28</sup> Under this alternative, one would also need to assume either: 1) that the regulator also constrains the level of capital (to the personally optimal, but economically inefficient, level), or 2) that the regulator fails to achieve his or her optimum. This can be seen by observing that if the regulator were assumed merely to constrain  $p_1$  and  $p_2$  (and allow the firm to choose  $k$  otherwise unfettered), the firm would take prices as given and thus (as in the theory of the competitive firm) minimize cost. But this outcome is inconsistent with utility maximization for the regulator.

<sup>26</sup>Note that at the efficient level of capital a decrease in  $k$  would result in increasing pressure from both sources.

<sup>27</sup>Note that this general approach to deriving regulatory rules can also provide theoretical rationales for other forms of regulation. For example, if (1) were restricted so that  $\partial u/\partial\pi = \partial u/\partial s = 0$ , the regulator would minimize  $p_1$  subject to  $\pi \geq 0$ . This case corresponds to a form of regulation called “residual pricing” and seems especially relevant to state regulation of telephone prices. (See Noll, 1986.)

<sup>28</sup>The AJ model has often been criticized because regulation results in prices which stay fixed for extended periods of time, not fixed rates of return. (See, for example, Klevorick, 1973, and Joskow, 1973, 1974.)

Thus the procedural emphasis on  $s$  and the regulator's resulting desire to keep  $s$  low (other things equal) is viewed as leading the regulator to desire some degree of economic inefficiency. Recalling the intuition discussed above, an overcapitalization result is anticipated for a variety of objective functions for the regulator. Apparently, such a result would obtain whenever the regulator is sufficiently concerned about  $s$  (at the cost-minimizing level of  $k$ , given prices however determined) to be willing to trade reduced profits for a lower rate of return.

Clearly various types of strategic behavior may be very important. Here we briefly consider values of  $p_1$  and  $s$  other than those that solve RCC. As with most instances of small numbers of agents, various outcomes are plausible, but:

**PROPOSITION 2:** *All Pareto optima between the regulator and the firm involve overcapitalization.*

The proposition emphasizes that political efficiency (narrowly defined) requires economic inefficiency. It can be seen using Figure 1.

For any  $p_1, p_1^0$  say, and  $s$  the firm and the regulator desire maximal profits and hence all Pareto optima must lie on  $\pi^*(p_1^0, s)$ . All points on this curve involving  $s < s_1^0$  are Pareto dominated by  $A$ . Points on  $\pi^*(p_1^0, s)$  to the right of  $B$  are Pareto dominated by point  $B$ . Thus, with  $p_1 = p_1^0$ , the segment of  $\pi^*(p_1^0, s)$  from  $A$  to  $B$  is a contract curve for the regulator and the firm. This logic applies for each value of  $p_1$  that would constrain the firm. Use of Corollary 1 to Proposition 1 then establishes Proposition 2.

#### IV. Uncertainty

While the analysis of the regulated firm under certainty is of substantial interest, uncertainty and dynamics seem essential to satisfactory theorizing in this area. In this section we consider uncertainty using the approach of Leland (1972), which is not explicitly dynamic but involves an aspect of timing (i.e., the *ex ante*, *ex post* distinction).

### A. The Two Approaches to Modeling the Regulated Firm Under Uncertainty

Most of the formal literature on the regulated firm under uncertainty (with symmetric information) has employed the device of regulatory constraint(s) as the sole representation of the role of the regulator. A leading exception is Klevorick (1973), which emphasizes procedural issues and explicitly introduces dynamic elements as well as various aspects of uncertainty. This impressive analysis, however, is complex<sup>29</sup> and does not provide a definite prediction about an issue of central concern: the efficiency of the input choices of the regulated firm.<sup>30</sup> The tractability difficulties encountered by Klevorick (1973) provide motivation for fortifying the regulatory-constraint approach, rather than abandoning<sup>31</sup> it in favor of more complex, process-oriented models. We focus on the regulatory constraint approach while attempting to address Klevorick's (1973, p. 58) central concern: "...it is fair to ask at what price the A-J model's results have come. In compressing reality into a tractable model, have important aspects of the regulatory process been suppressed, and hence, their theoretical and policy implications been lost? The answer appears to be in the affirmative." We study the motives of the regulator, taking into account important aspects of the regulatory process, and examine the condi-

tions under which the role of the regulator can be usefully represented by (one or more) regulatory constraints.

We attempt to consider all important sources of uncertainty and allow each source of uncertainty to enter in a fairly general way. Both cost and demand uncertainty are considered, paying particular attention to an aspect of cost uncertainty that seems especially important in the case of regulated firms: the cost of capital.<sup>32</sup>

### B. The Regulator's Choice of Regulatory Constraints Under Uncertainty

Let  $\omega \in \Omega$  index states of the world and  $f(\omega)$  govern the realization of  $\omega$ . Let  $l$  and  $k$  denote a labor input and the capital input, both of which are *ex ante* controls. Denote by  $x$  the vector of *ex post* controls, which represent materials and freely variable types of labor. Let  $r(\omega)$ ,  $w(\omega)$ , and  $v(\omega)$  denote respectively the prices of  $k$ ,  $l$ , and  $x$ .

Demands for the two goods, which must be satisfied *ex post*, are written as

$$(8a) \quad q_1^d = q_1(p_1, \omega), \quad \text{and}$$

$$(8b) \quad q_2^d = q_2(p_2, \omega).$$

The output prices  $p_1$  and  $p_2$  are assumed fixed *ex ante*, to reflect an institutional fact that many have argued is critical.<sup>33</sup> Denote by  $q_1$  and  $q_2$  the production levels of the two goods, and use the market-clearing assumption to write the minimum cost of

<sup>29</sup>Evans and Garber (1985, fn. 28) attempt a concise description of Klevorick's model. See also William T. Ziemba (1974) and Klevorick (1974).

<sup>30</sup>Unambiguous predictions require special assumptions about the production technology. Klevorick (1973, p. 78) argues for the plausibility of a set of conditions under which the regulated firm will use *less* capital than minimizes the cost of its output.

<sup>31</sup>We are aware of three other formal, positive analyses of the regulated firm under (symmetric information) uncertainty that fall outside the "regulatory constraint" tradition: Burness, Montgomery, and Quirk, 1980, §2; Bawa and Sibley, 1980; and Braeutigam and Quirk, 1984. The first assumes uncertainty about only the cost of capital and focuses on an interesting, but very special issue. The second considers uncertainty about only the timing of future rate reviews. The third considers demand uncertainty only and results in ambiguous predictions.

<sup>32</sup>The literature on the unregulated firm under uncertainty generally abstracts from inflation, but when fixed nominal prices are an important feature of the problem, as is the case with public utilities, uncertainty about the (nominal) cost of capital due to uncertainty about future rates of inflation seems central. In the formal positive literature, Burness, Montgomery, and Quirk (1980) emphasize this aspect of uncertainty; assuming, in fact, that the price of capital is the only uncertain quantity.

<sup>33</sup>See, for example, Klevorick, 1973; Joskow, 1973, 1974. Here we allow the regulator to set prices that stay fixed no matter what state of the world is realized (one directly and the other indirectly); but these prices take into account the political and economic ramifications for the regulator.

*ex post* inputs (given  $k$  and  $l$ ) as

$$(9) \quad c = c(q_1(p_1, \omega), q_2(p_2, \omega), v(\omega), k, l).$$

We assume throughout that both the regulator and the firm know the cost and demand structure and share knowledge of  $f(\omega)$ .

Denote by  $z \equiv (p_1, p_2, l, k)$  the vector of *ex ante* controls and by  $x(z, \omega)$  the cost-minimizing choice of *ex post* inputs given  $z$  and  $\omega$ . Then profits depend on the realization of  $\omega$  and are written as

$$(10) \quad \begin{aligned} \pi(z, \omega) &= \pi(z, x(z, \omega), \omega) \\ &= p_1 q_1(p_1, \omega) + p_2 q_2(p_2, \omega) \\ &\quad - c(q_1(p_1, \omega), q_2(p_2, \omega), v(\omega), k, l) \\ &\quad - r(\omega)k - w(\omega)l. \end{aligned}$$

We assume throughout that  $\pi(z, \omega)$  is strictly concave in  $z$ . In addition, the rate of return is, in general, subject to uncertainty

$$(11) \quad s(z, \omega) = \pi(z, \omega)/k + r(\omega).$$

Note from (10), however, that if  $r(\omega)$  is the only uncertain quantity  $s(z, \omega) = s(z)$  is nonstochastic.

In the case of certainty the firm was assumed to maximize  $\pi$ . With uncertainty, we assume that the firm evaluates states of the world according to the utility function

$$\tilde{u} = \tilde{u}(\pi(z, \omega)),$$

with  $\tilde{u}' > 0$  and  $\tilde{u}'' \leq 0$ , and that the firm seeks *ex ante* to maximize expected<sup>34</sup> utility. Similarly, to examine uncertainty we assume that advocates for the buyers of both goods evaluate  $s(z, \omega)$  by means of

$$\hat{u} = \hat{u}(s(z, \omega)),$$

with  $\hat{u}' < 0$  and  $\hat{u}'' \leq 0$ . These generalizations require, in turn, a generalization of the regulator's objective (1).

The specification of (1) was based in part on the assumption that pressure exerted by the managers of the firm and its stockholders depends on profits, their objective in the certainty case. The advocates for buyers of both goods were assumed to focus on the rate of return. Here we assume that political pressure by these groups is determined by the utilities associated with the (*ex post*) realizations of  $\pi$  and  $s$ , respectively.<sup>35</sup> Thus, using the logic of Section II, we arrive at

$$(1') \quad u = u(\tilde{u}(\pi(z, \omega)), \hat{u}(s(z, \omega)), -p_1)$$

as the objective function for our regulator in the case of uncertainty. The legal requirement that the firm must be able to raise capital (without confiscation of equity-holders' property) is that  $E\pi(z, \omega) \geq 0$ , or equivalently

$$(12) \quad Es(z, \omega) \geq Er(\omega).$$

We assume throughout that (12) is satisfied as a strict inequality.

Our representation of uncertainty is quite general and it seems unwise to restrict further the model in this dimension. In order to obtain definite results, we assume that (1') is additively separable, and that the regulator is risk neutral with respect to the utility level

<sup>35</sup>This assumption seems most reasonable, but it is not unassailable. In some settings (for example, in the presence of very sophisticated political actors) pressure might be more reasonably modeled as resulting from the (*ex ante*) distributions of  $\pi$  and  $s$ , which, by assumption, are known by all *ex ante*. In effect, we assume that (like the members of the pressure groups) the regulator's fortunes depend on the realized state of the world, rather than assuming that regulators do not have a stake in the random forces that are important to these groups. Thus if pressure were assumed to be exerted *ex ante*, the regulator would face no uncertainty. (Formally, assuming, for example, that the buyers' group seeks to maximize  $E\hat{u}$ , the arguments of the regulator's objective function would be  $E\tilde{u}$ ,  $E\hat{u}$ , and  $p_1$ , none of which depends on  $\omega$ .)

<sup>34</sup>All expectations are taken over  $\Omega$  using  $f(\omega)$ .

of the firm:

$$(13) \quad u(\tilde{u}(\pi(z, \omega)), \hat{u}(s(z, \omega)), -p_1) \\ = \tilde{u}(\pi(z, \omega)) + W(\hat{u}(s(z, \omega)), -p_1).$$

The separability and risk-neutrality assumptions are expected to be useful approximations in a broad range of circumstances.

As indicated above, regulation results in prices which stay fixed over time, not fixed rates of return. Thus  $p_1$  and  $p_2$  are taken to be *ex ante* controls. We take this into account in a way that preserves another feature of regulation which is believed to be important: the ability of the firm and other groups to influence the regulator's choice of prices. In particular, the regulator is assumed to constrain  $p_1$  directly, and the rest of the revenue constraint on the firm is represented indirectly by a constraint involving the firm's rate of return on capital. Two such constraints involving  $s(z, \omega)$  are considered.

The first constraint involving  $s(z, \omega)$  was used in the earliest formal papers extending the regulatory constraint approach to the case of uncertainty: Peles and Stein, 1976; and Perrakis, 1976a, b. In the spirit of AJ's ceiling on the rate of return, these papers specify an *ex post* regulatory constraint on the maximal realization of  $s(z, \omega)$ , censoring<sup>36</sup> its distribution from above,

$$(14) \quad s(z, \omega) \leq \tilde{s} \quad \forall \omega \in \Omega,$$

where  $\tilde{s}$  is a constant, presumably chosen by the regulator. Perrakis, 1976b, p. 415, refers to this regulatory constraint as a "natural" extension of AJ, and this representation has also been used by Perrakis, 1983; Rau, 1979; Peles and Stein, 1979; and Das, 1980.<sup>37</sup> This constraint can be criticized for lack of any

means by which it could be enforced at a cost that the regulator is willing to bear (i.e., how is the rate of return equated to  $\tilde{s}$  if it would otherwise exceed  $\tilde{s}$ ?). Our critique, however, centers on a distinct issue: Suppose (14) could be perfectly and costlessly enforced; would it serve the interests of the regulator to impose it?

Constraint (14) has been adopted in the literature without any apparent consideration of the objectives and opportunities of the regulator. It is a "natural" extension of the AJ constraint, but it is hardly *the* natural extension. Consider as an alternative constraining *ex ante* the mathematical expectation of the rate of return<sup>38</sup>

$$(15) \quad Es(z, \omega) \leq \delta,$$

where  $\delta$  is some constant exceeding  $Er(\omega)$ . We now show that constraints of the form (15) can serve the interests of the regulator, at least in some plausible special cases, while constraints like (14) cannot.<sup>39</sup>

<sup>38</sup>This type of constraint was used by R. Chapman and L. Waverman (1979), who motivate (pp. 109-110) it in terms of a practical concern: the potential impossibility of meeting demand and a rate of return constraint *ex post*. The view that price regulation constrains the expectation of  $s(z, \omega)$  *ex ante* seems most appropriate in the case of "forward-looking" regulation, which is the focus of the theoretical literature. See, for example, Sibley and Bailey (1978) who contrast "forward-looking" and "myopic" regulation.

<sup>39</sup>In addition to the perspective emphasized in the text, constraint (15) may also be preferred to (14) because it is responsive to some major criticisms of the AJ model. (See, for example, Joskow, 1974.) First, (15) takes into account that regulated prices stay fixed for periods of time, not rates of return: we view (15) as representing a result of fixed prices. Constraint (14) cannot be viewed in this way because satisfying (14) may require a price adjustment once  $\omega$  is realized. Second, (15) takes account of the fact that regulators often have neither the desire nor the resources to monitor continually the fortunes of the firms they regulate: constraint (14) could be enforced only by the costly means of frequent monitoring and, when warranted, negotiation of price decreases or additional formal rate reviews. See, for example, Joskow, 1972a, ch. II, for a description of these regulatory processes in New York State. Implementation of constraint (15) does not require regulatory action *ex post*.

<sup>36</sup>That is, if  $\omega$  is such that  $s(z, \omega)$  would exceed  $\tilde{s}$ , then it is assumed that the realized rate of return is  $\tilde{s}$ .

<sup>37</sup>Burness, Montgomery, and Quirk (1980, §1) specify a constraint slightly different from (14): in place of the actual value of the capital stock (which appears in the denominator of  $s(z, \omega)$  and is stochastic because the price of capital is uncertain) they substitute its mathematical expectation.



To discover the circumstances under which constraints of the forms of (14) and (15) can give the regulator complete control, assume that the regulator is an expected utility maximizer constrained by (12) and consider the regulator's choices of the *ex ante* controls if that person had complete control. The relevant optimization problem is then<sup>40</sup>

$$\text{RCCU: } \max E[\tilde{u}(\pi(z, \omega)) + W(\hat{u}(s(z, \omega)), -p_1)]$$

$$\text{subject to } Es(z, \omega) \geq Er(\omega).$$

Denote the solution to RCCU as  $z^* \equiv (p_1^*, p_2^*, l^*, k^*)$ . We seek conditions under which the regulator can induce the firm to choose  $p_2^*$ ,  $l^*$ , and  $k^*$  merely by imposing the constraint  $p_1 = p_1^*$  and a constraint of the form (14) or (15).

First, a constraint of the form (14) can result in the same solution as RCCU, namely  $z^*$ , only if this constraint is satisfied by  $z^*$ . Thus consider the constraint

$$(14') \quad s(z, \omega) \leq \bar{s} \equiv \max_{\omega} \{s(z^*, \omega); \omega \in \Omega\},$$

so that  $z^*$  is feasible if this constraint is imposed on the firm. Similarly, a constraint on the expected rate of return can be satisfied by  $z^*$  only if

$$(15') \quad Es(z, \omega) = \bar{s} \equiv Es(z^*, \omega).^{41}$$

Assume that the firm chooses  $p_2$ ,  $l$ , and  $k$  to maximize its expected utility, and consider the firm's problems imposing as alternatives

$$(14') \text{ and } (15')^{42}$$

$$\text{MCMU: } \max_{p_2, l, k} E\tilde{u}(\pi(z, \omega))$$

$$\text{subject to } p_1 = p_1^* \text{ and } s(z, \omega) \leq \bar{s},$$

and

$$\text{ECMU: } \max_{p_2, l, k} E\tilde{u}(\pi(z, \omega))$$

$$\text{subject to } p_1 = p_1^* \text{ and } Es(z, \omega) = \bar{s}.$$

Denote by  $\tilde{z}$  and  $\bar{z}$  the respective solutions to MCMU and ECMU, and assume that  $z^*$ ,  $\tilde{z}$ , and  $\bar{z}$  are unique.<sup>43</sup>

Define  $V(\hat{u}(s(z, \omega))) \equiv W(\hat{u}(s(z, \omega)), -p_1^*)$ . In Appendix B we prove

**PROPOSITION 3:** (i) *The choice vectors  $\tilde{z}$  and  $z^*$  coincide if and only if  $EV(\hat{u}(s(\tilde{z}, \omega))) = EV(\hat{u}(s(z^*, \omega)))$ , and (ii) *The choice vectors  $\bar{z}$  and  $z^*$  coincide if and only if  $EV(\hat{u}(s(\bar{z}, \omega))) = EV(\hat{u}(s(z^*, \omega)))$ .**

Proposition 3(i) and 3(ii) give the necessary and sufficient conditions for the regulator to exercise complete control using, respectively, (14') or (15') to constrain the firm's rate of return. We now consider the economics of the provisos involving the expectations of  $V(\hat{u}(s(z, \omega)))$  evaluated at  $z^*$ ,  $\tilde{z}$ , and  $\bar{z}$ .

In view of (15'), it is obvious that if  $V(\cdot)$  and  $\hat{u}(\cdot)$  were linear, constraint (15') guarantees that the proviso of Proposition 3(ii) is satisfied. Thus if the regulator were risk neutral with respect to the utility of the advocates for the buyers of both goods, and these advocates were risk neutral with respect to the rate of return, the regulator can achieve

<sup>40</sup>RCCU denotes "regulator has complete control-uncertainty."

<sup>41</sup>In order to achieve complete control, the regulator might specify an equality constraint (as in (15')) or an upper bound on  $Es(z, \omega)$  (as in (15)). The formulation in (15') merely employs the assumption that the appropriate inequality constraint will constrain the firm, or that the regulator needs to constrain  $Es$  in addition to  $p_1$  in order to achieve complete control.

<sup>42</sup>MCMU denotes "maximal-realization-constrained monopoly, uncertainty" and ECMU denotes "expected-realization-constrained monopoly, uncertainty."

<sup>43</sup>The uniqueness assumptions are made solely for expositional convenience. The solutions to RCCU, MCMU, and ECMU can also be viewed as sets of choices of  $z$  without materially influencing what follows; see the footnote in Appendix B.

complete control by constraining  $Es(z, \omega)$ . But linearity of  $V(\cdot)$  and  $\hat{u}(\cdot)$  would allow complete control by constraining the maximal realization of  $s(z, \omega)$  only in *very* special circumstances: Requiring the maximal realization of  $s(\bar{z}, \omega)$  to coincide with that of  $s(z^*, \omega)$  will generally not require their means to coincide.<sup>44</sup>

It is also obvious that the provisos of Proposition 3(i) and 3(ii) would be satisfied if the distributions of  $s(\bar{z}, \omega)$  and  $s(\bar{z}, \omega)$ , respectively, coincide with the distribution of  $s(z^*, \omega)$ . Here a (statistically) degenerate, but economically interesting, special case becomes relevant. Recall that if  $r(\omega)$  is the only source of uncertainty then  $s(z, \omega)$  is nonstochastic. Thus if the cost of capital is the sole source of uncertainty, imposition of either (14') or (15') could give the regulator complete control.<sup>45</sup> We note, however, that in none of the papers specifying a regulatory constraint of the form (14) is the rate of return, in fact, nonstochastic.

Finally, it is also clear that the proviso in Proposition 3(ii) would be satisfied if uncertainty about the rate of return is introduced in such a way that the mean of  $s(z, \omega)$  completely characterizes its distribution. This possibility seems uninteresting economically, but such specifications have, in fact, had an impact on the development of the literature.<sup>46</sup> And, as in the case where  $V(\cdot)$  and  $\hat{u}(\cdot)$  are linear, even in this case the regulator can generally achieve complete control using (14')

only in very special circumstances: once again, constraining the maximal realizations of  $s(z^*, \omega)$  and  $s(\bar{z}, \omega)$  to coincide will generally not cause their means to coincide.

We summarize this discussion by

**PROPOSITION 4:** *If the behavior of the firm is to maximize the expected utility of profits subject to the constraints imposed by the regulator:*

(i) *A constraint on the expected rate of return can give the regulator complete control if the regulator and the advocates who focus on the rate of return are risk neutral. A constraint on the maximal realization of the rate of return will, in general, not give the regulator complete control in this case.*

(ii) *Either rate of return constraint can give the regulator complete control when the cost of capital is the only source of uncertainty.*

(iii) *A constraint on the expected rate of return can give the regulator complete control under an economically uninteresting specification which has received considerable attention. A constraint on the maximal realization of the rate of return generally fails to give the regulator complete control in this case as well.*

Thus we conclude that in some special, but economically interesting, circumstances the regulator would be perfectly willing to impose a constraint of the form of (15') but not a constraint of the form of (14'), even assuming costless enforcement. Thus we have a theoretical rationale for "expected rate-of-return" regulation, but not "realized rate-of-return" regulation. Moreover, consideration of the relative costs to the regulator of enforcing the constraints works toward the same conclusion. From the point of view of the regulator, constraining the expected rate of return seems preferable to constraining the maximal realization of the rate of return. We conclude that the regulatory-constraint approach (with symmetric uncertainty) is better pursued using a constraint on the expected rate of return than on the realized rate of return.

Thus we have established a basis for modeling the firm as subject to an *ex ante* constraint on the expected value of  $s$  rather than an *ex post* constraint on the realization

<sup>44</sup> Recall that in the relevant literature, the distribution of  $s(\bar{z}, \omega)$  is viewed as a censored distribution, while the distribution of  $s(z^*, \omega)$  is not.

<sup>45</sup> Since  $s$  is nonstochastic, specifying either the maximal realization or the mean completely characterizes its (degenerate) distribution.

<sup>46</sup> Peles and Stein, 1976, introduce uncertainty about  $s$  directly (rather than through components of demand or cost) in both an additive or multiplicative fashion:

$$s(z, \omega) = s(z) + \varepsilon(\omega), \text{ or } s(z, \omega) = s(z)(1 + \varepsilon(\omega)),$$

where  $s(z)$  is constant over  $\omega$ ,  $\varepsilon$  has a zero mean, and  $\varepsilon$  is independent of  $z$ . Then with either specification  $Es(z, \omega) = s(z)$  follows directly. In these specifications for  $s(z, \omega)$  the mean of  $s$  completely determines its distribution since  $\varepsilon(\omega)$  is independent of the *ex ante* choices  $z$ .

of  $s$ . But an important question remains: is such a formulation tractable? In Evans and Garber (1985) we present analyses of ECMU which suggest an affirmative answer. There we analyze two special cases for which the regulator has complete control. In each case we focus on the issue: will the regulated firm overcapitalize? We briefly report our conclusions, referring interested readers to Evans and Garber (1985) for details and proofs.

First we consider the case in which  $V(\hat{u}(s(z, \omega)))$  is linear in  $s(z, \omega)$ , the sources of uncertainty are unrestricted, and the firm maximizes expected profits. The following proposition summarizes our conclusion:

**PROPOSITION 5:** *Under our general specification of uncertainty, in the case of full risk neutrality (i.e., on the parts of the regulator, the firm, and the advocates focusing on the rate of return), the optimally regulated firm uses an inefficiently large level of capital for the ex ante distribution of outputs which it faces.*

Next we consider the case in which  $r(\omega)$  is the only source of uncertainty, but we allow the functions  $\hat{u}(\pi(z, \omega))$  and  $V(\hat{u}(s(z, \omega)))$  to exhibit risk aversion. Intuitively, with risk aversion and the cost of capital as the only variable subject to uncertainty, we have incorporated forces which tend to deter the firm from overcapitalizing, because large capital stocks must increase the riskiness associated with the distribution of profits when the cost of capital is the only uncertain quantity. Thus, as one would expect, an ambiguous result emerges:

**PROPOSITION 6:** *With the cost of capital as the only source of uncertainty, in the case of risk aversion the optimally regulated firm may choose an inefficiently large or inefficiently small level of the capital input. The firm is more likely to overcapitalize: (a) the less risk averse is the firm (in the Arrow-Pratt sense), and (b) the smaller is the variance of  $r(\omega)$ .*

### V. Concluding Comments

Understanding the source of regulatory rules is an interesting and important issue in

its own right. It is also a first step toward predicting the behavior of regulated firms. The "regulatory-constraint" approach to predicting this behavior is attractive. One reason is the possibility that useful models may be analyzed straightforwardly by employing standard optimization tools. But specification of useful constraints or regulatory rules seems to require explicit consideration of the interests and opportunities of the regulator. We have shown that constraints that serve the regulator can be deduced using an objective function for the regulator based on consideration of various human concerns, and various aspects of the politics and processes of public-utility regulation.

The analysis also provides theoretical rationales for rate-of-return regulation in the certainty case, and a particular type of rate-of-return regulation under symmetric uncertainty. Other forms of regulation should also be explicable employing the basic approach. An important, and apparently very challenging, extension would involve combining objectives for "human" regulators with asymmetric information, which also seems to be a central feature of most regulatory environments.

The fact that much of the literature on the regulated firm under uncertainty has employed a constraint with little, if any, a priori appeal underscores the potential value of deducing regulatory constraints rather than assuming them. But much remains to be done, even in the nature of general theorizing. Valuable extensions would involve consideration of other objective functions and dynamics, as well as asymmetric information.

### APPENDIX A: PROOFS OF THE COROLLARIES TO PROPOSITION 1

Consider first the subsidiary problem which is (implicitly) solved in RCC:

$$(5) \quad \pi^*(p_1, s) = \max_{p_2, k} \pi(p_1, p_2, k)$$

subject to  $sk = \pi + rk$ ,

with associated Lagrangian

$$\mathcal{L} = \pi(p_1, p_2, k) + \lambda(sk - \pi - rk).$$

The necessary conditions are

$$(A1a) \quad (1 - \lambda)\pi_{p_2} = 0,$$

$$(A1b) \quad (1 - \lambda)\pi_k = -\lambda(s - r), \quad \text{and}$$

$$(A1c) \quad sk - \pi - rk = 0,$$

where subscripts denote partial derivatives. Assuming that  $s > r$  and  $\lambda > 0$  (i.e., that the rate of return constraint is binding), (A1b) implies that  $(1 - \lambda)\pi_k \neq 0$  and thus that  $(1 - \lambda) \neq 0$ . Since  $(1 - \lambda) \neq 0$ , (A1a) then implies that  $\pi_{p_2} = 0$ . The second-order conditions and the concavity of  $\pi(p_1, p_2, k)$  then imply  $1 - \lambda > 0$ . Thus we have Corollary 2 to Proposition 1:  $0 < \lambda < 1$ . Finally, note that  $0 < \lambda < 1$  and  $s > r$  imply, using (A1b), that  $\pi_k < 0$ . Corollaries 1, 3, and 4 relate to the comparative statics of the problem given by (5). Total differentiation of equations (A1) yields

$$(A2) \quad \begin{bmatrix} (1 - \lambda)\pi_{p_2 p_2} & (1 - \lambda)\pi_{p_2 k} & 0 \\ (1 - \lambda)\pi_{p_2 k} & (1 - \lambda)\pi_{kk} & s - r - \pi_k \\ 0 & s - r - \pi_k & 0 \end{bmatrix} \times \begin{bmatrix} dp_2 \\ dk \\ d\lambda \end{bmatrix} = \begin{bmatrix} 0 & -(1 - \lambda)\pi_{p_2 p_1} \\ -\lambda & -(1 - \lambda)\pi_{kp_1} \\ -k & \pi_{p_1} \end{bmatrix} \begin{bmatrix} ds \\ dp_1 \end{bmatrix}.$$

To establish Corollary 3, use (A2) to compute

$$\frac{\partial k}{\partial s} = (s - r - \pi_k)k(1 - \lambda)\pi_{p_2 p_2}/|H|,$$

where  $|H|$  is the determinant of the bordered Hessian matrix on the left-hand side (LHS) of (A2). Using  $s > r$ ,  $\pi_k < 0$ ,  $(1 - \lambda) > 0$ ,  $\pi_{p_2 p_2} < 0$  (by the concavity of  $\pi(p_1, p_2, k)$ ) and the second-order condition of (5) that  $|H| > 0$ , we conclude  $\partial k/\partial s < 0$ . Thus Corollary 3 to Proposition 1 is established.

To verify Corollary 4, use (A2) to compute

$$\frac{\partial \lambda}{\partial s} = (s - r - \pi_k)\lambda(1 - \lambda)\pi_{p_2 p_2}/|H|,$$

which is negative.

Finally, to see that the firm uses too much capital to minimize the production cost of its output bundle hold  $p_1$  and  $p_2$ , and thus the output mix, fixed and compute  $\partial k/\partial s = -k/(s - r - \pi_k)$ . This derivative is negative as long as the rate of return constraint is binding, and thus we have a global result. In particular, consider a firm which faces the same prices (and, thus, output combination) as the regulated firm, but is otherwise unconstrained. This firm uses the efficient level of  $k$ . Now, introduce a constraint on  $s$  and tighten this constraint (holding  $p_1$  and  $p_2$ , and thus  $q_1$  and  $q_2$ , fixed) until the ceiling on  $s$  equals the optimal level of  $s$  in RCC. Since  $\partial k/\partial s < 0$  whenever  $\lambda > 0$ , we have established that the regulated firm uses more than the efficient level of capital, as stated in Corollary 1 to Proposition 1.

#### APPENDIX B: PROOF OF PROPOSITION 3

Proposition 3(ii) is proved in detail. The proof of Proposition 3(i) follows directly, *mutatis mutandis*.

First we prove sufficiency (i.e., that  $EV(*) \equiv EV(\hat{u}(s(z^*, \omega))) = EV(\hat{u}(s(\bar{z}, \omega))) \equiv EV(-)$  implies  $z^* = \bar{z}$ ). Since  $z^*$  solves RCCU (i.e., maximizes  $Eu(z, \omega)$ ), we have by (13):

$$(B1) \quad E\bar{u}(\pi(z^*, \omega)) + EV(\hat{u}(s(z^*, \omega))) \geq E\bar{u}(\pi(\bar{z}, \omega)) + EV(\hat{u}(s(\bar{z}, \omega))).$$

The proviso of Proposition 3(ii),  $EV(*) = EV(-)$ , (B.1), and subtraction then imply

$$(B2) \quad E\bar{u}(\pi(z^*, \omega)) \geq E\bar{u}(\pi(\bar{z}, \omega)).$$

But, since  $\bar{z}$  solves ECMU and  $z^*$  is feasible in ECMU, we must also have

$$(B3) \quad E\bar{u}(\pi(\bar{z}, \omega)) \geq E\bar{u}(\pi(z^*, \omega)),$$

and thus by (B2) and (B3)

$$(B4) \quad E\tilde{u}(\pi(\bar{z}, \omega)) = E\tilde{u}(\pi(z^*, \omega)).$$

Use of the proviso and (B4) then yields

$$\begin{aligned} (B5) \quad E u(z^*, \omega) &= E\tilde{u}(\pi(z^*, \omega)) + EV(\hat{u}(s(z^*, \omega))) \\ &= E\tilde{u}(\pi(\bar{z}, \omega)) + EV(\hat{u}(s(\bar{z}, \omega))) \\ &= E u(\bar{z}, \omega), \end{aligned}$$

which establishes that the regulator gets equal expected utility from  $z^*$  (the solution to RCCU) and  $\bar{z}$  (the solution to ECMU).<sup>47</sup> The assumption that  $z^*$  is unique then implies  $z^* = \bar{z}$ . Thus the sufficiency of the proviso is established. Necessity of the proviso follows trivially since  $z^* = \bar{z}$  implies immediately  $EV(\hat{u}(s(z^*, \omega))) = EV(\hat{u}(s(\bar{z}, \omega)))$ .

Proposition 3(i) can be proved using the same logic merely substituting  $\bar{z}$  for  $\bar{z}$  and MCMU for ECMU in the above. (In essence, the proofs coincide because neither proof uses the form of the rate of return constraint.)

<sup>47</sup>Note that the uniqueness of  $z^*$  and  $\bar{z}$  has not to this point been invoked, and thus we get the essence of complete control, namely indifference on the part of the regulator, without the uniqueness assumptions.

## REFERENCES

- Averch, Harvey and Johnson, Leland L., "Behavior of the Firm Under Regulatory Constraint," *American Economic Review*, December 1962, 52, 1052-69.
- Bailey, Elizabeth E., *Economic Theory of the Regulatory Constraint*, Lexington, MA: Lexington Books, 1973.
- \_\_\_\_\_ and Coleman, Roger D., "The Effect of Lagged Regulation in an Averch-Johnson Model," *Bell Journal of Economics and Management Science*, Spring 1971, 2, 278-92.
- \_\_\_\_\_ and Malone, John C., "Resource Allocation and the Regulated Firm," *Bell Journal of Economics and Management Science*, Spring 1970, 1, 129-42.
- Baron, David P. and Besanko, David, "Regulation, Asymmetric Information, and Auditing," *The Rand Journal of Economics*, Winter 1984, 15, 447-70.
- \_\_\_\_\_ and DeBondt, Raymond R., "On the Design of Regulatory Price Adjustment Mechanisms," *Journal of Economic Theory*, February 1981, 24, 70-94.
- \_\_\_\_\_ and Myerson, Roger B., "Regulating a Monopolist with Unknown Costs," *Econometrica*, July 1982, 50, 911-30.
- Baumol, William J. and Klevorick, Alvin K., "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion," *Bell Journal of Economics and Management Science*, Autumn 1970, 1, 162-90.
- Bawa, Vijay S. and Sibley, David S., "Dynamic Behavior of a Firm Subject to Stochastic Regulatory Reviews," *International Economic Review*, October 1980, 21, 627-42.
- Becker, Gary S., "A Theory of Competition Among Pressure Groups for Political Influence," *Quarterly Journal of Economics*, August 1983, 98, 371-400.
- Braeutigam, Ronald R. and Quirk, James P., "Demand Uncertainty and the Regulated Firm," *International Economic Review*, February 1984, 25, 45-60.
- Breyer, Stephen, *Regulation and its Reform*, Cambridge, MA: Harvard University Press, 1982.
- Brown, Gardner, Jr., and Johnson, M. Bruce, "Public Utility Pricing and Output Under Risk," *American Economic Review*, March 1969, 59, 119-29.
- Burness, H. Stuart, Montgomery, W. David and Quirk, James P., "Capital Contracting and the Regulated Firm," *American Economic Review*, June 1980, 70, 342-54.
- Chapman, R. and Waverman, L., "Risk Aversion, Uncertain Demand and the Effects of a Regulatory Constraint," *Journal of Public Economics*, February 1979, 11, 107-21.
- Corey, Gordon R., "The Averch-Johnson Proposition: A Critical Analysis," *Bell Jour-*

- nal of Economics and Management Science*, Spring 1971, 2, 358-73.
- Das, Satya P., "On the Effect of Rate of Return Regulation Under Uncertainty," *American Economic Review*, June 1980, 70, 456-60.
- Davis, E. G., "A Dynamic Model of the Regulated Firm with a Price Adjustment Mechanism," *Bell Journal of Economics and Management Science*, Spring 1973, 4, 270-82.
- Demski, Joel S. and Sappington, David E. M., "Hierarchical Regulatory Control," *Rand Journal of Economics*, Autumn 1987, 18, 369-83.
- Eckert, Ross D., "On the Incentives of Regulators: The Case of Taxicabs," *Public Choice*, Spring 1973, 14, 83-99.
- \_\_\_\_\_, "The Life Cycle of Regulatory Commissioners," *Journal of Law and Economics*, April 1981, 24, 113-20.
- Evans, Lewis and Garber, Steven, "Public Utility Regulation: A Theory of Constraints that Serve the Regulator," unpublished manuscript, January 4, 1985, Carnegie-Mellon University.
- Finsinger, Jörg and Vogelsang, Ingo, "Alternative Institutional Frameworks for Price Incentive Mechanisms," *Kyklos*, 1981, 34, 388-404.
- Gormley, William T., Jr., *The Politics of Public Utility Regulation*, Pittsburgh, PA: University of Pittsburgh Press, 1983.
- Hagerman, Robert L. and Ratchford, Brian T., "Some Determinants of Allowed Rates of Return on Equity to Electric Utilities," *Bell Journal of Economics*, Spring 1978, 9, 46-55.
- Joskow, Paul L., (1972a) *A Behavioral Theory of Public Utility Regulation*, unpublished doctoral dissertation, Yale University, 1972.
- \_\_\_\_\_, (1972b) "The Determination of the Allowed Rate of Return in a Formal Regulatory Hearing," *Bell Journal of Economics and Management Science*, Autumn 1972, 3, 632-44.
- \_\_\_\_\_, "Pricing Decisions of Regulated Firms: A Behavioral Approach," *Bell Journal of Economics and Management Science*, Spring 1973, 4, 118-40.
- \_\_\_\_\_, "Inflation and Environmental Concern: Structural Change in the Process of Public Utility Regulation," *Journal of Law and Economics*, October 1974, 17, 291-327.
- \_\_\_\_\_, and Noll, Roger G., "Regulation in Theory and Practice: An Overview," in Gary Fromm, ed., *Studies in Public Regulation*, Cambridge, MA: MIT Press, 1981, ch. 1.
- Kahn, Alfred E., *The Economics of Regulation*, Vol. I, New York: Wiley & Sons, 1970.
- Klevorick, Alvin K., "The Graduated Fair Return: A Regulatory Proposal," *American Economic Review*, June 1966, 56, 477-84.
- \_\_\_\_\_, "The 'Optimal' Fair Rate of Return," *Bell Journal of Economics and Management Science*, Spring 1971, 2, 122-43.
- \_\_\_\_\_, "The Behavior of a Firm Subject to Stochastic Regulatory Review," *Bell Journal of Economics and Management Science*, Spring 1973, 4, 57-88.
- \_\_\_\_\_, "The Behavior of a Firm Subject to Stochastic Regulatory Review: Correction," *Bell Journal of Economics and Management Science*, Autumn 1974, 5, 713-14.
- Laffont, Jean-Jacques and Tirole, Jean, "Using Cost Observation to Regulate Firms," *Journal of Political Economy*, June 1986, 94, 614-41.
- Leland, Hayne E., "Theory of the Firm Facing Uncertain Demand," *American Economic Review*, June 1972, 62, 278-91.
- \_\_\_\_\_, "Regulation of Natural Monopolies and the Fair Rate of Return," *Bell Journal of Economics and Management Science*, Spring 1974, 5, 3-15.
- Lewis, Tracy R. and Sappington, David E. M., "Regulating a Monopolist with Unknown Demand," unpublished manuscript, January 1987.
- Littlechild, Stephen C., "A State Preference Approach to Public Utility Pricing and Investment Under Risk," *Bell Journal of Economics and Management Science*, Spring 1972, 3, 340-45.
- McNicol, David L., "The Comparative Statics Properties of the Theory of the Regulated Firm," *Bell Journal of Economics and Management Science*, Autumn 1973, 4, 428-53.

- Noll, Roger G., "Government Regulatory Behavior: A Multidisciplinary Survey and Synthesis," in Roger G. Noll, ed., *Regulatory Policy and the Social Sciences*, Berkeley, CA: University of California Press, 1985, ch. 2.
- \_\_\_\_\_, "State Regulatory Responses to Competition and Divestiture in the Telecommunications Industry," in Ronald E. Grieson, ed., *Antitrust and Regulation*, Lexington, MA: Lexington Books, 1986, ch. 9.
- Peles, Yoram C. and Stein, Jerome L., "The Effect of Rate of Return Regulation is Highly Sensitive to the Nature of the Uncertainty," *American Economic Review*, June 1976, 66, 278-89.
- \_\_\_\_\_, "On Regulation and Uncertainty: Reply," *American Economic Review*, March 1979, 69, 195-99.
- Peltzman, Sam, "Toward a More General Theory of Regulation," *Journal of Law and Economics*, August 1976, 19, 109-48.
- Perrakis, Stylianos, (1976a) "Rate of Return Regulation of a Monopoly Firm with Random Demand," *International Economic Review*, February 1976, 17, 149-62.
- \_\_\_\_\_, (1976b) "On the Regulated Price-Setting Monopoly Firm with a Random Demand Curve," *American Economic Review*, June 1976, 66, 410-16.
- \_\_\_\_\_, "The Value of the Firm Under Regulation and the Theory of the Firm Under Uncertainty: An Integrated Approach," in L. Courville et al., eds., *Economic Analysis of Telecommunications: Theory and Applications*, Amsterdam: North-Holland, 1983, 397-413.
- Phillips, Charles F., Jr., *The Regulation of Public Utilities*, Arlington, VA: Public Utility Reports, Inc., 1984.
- Posner, Richard A., "Taxation by Regulation," *Bell Journal of Economics and Management Science*, Spring 1971, 2, 22-50.
- Rau, Nicholas, "On Regulation and Uncertainty: Comment," *American Economic Review*, March 1979, 69, 190-94.
- Riordan, Michael H., "On Delegating Price Authority to a Regulated Firm," *Rand Journal of Economics*, Spring 1984, 15, 108-15.
- Roberts, R. Blaine, Maddala, G. S. and Enholm, Gregory, "Determinants of the Requested Rate of Return and the Rate of Return Granted in a Formal Regulatory Process," *Bell Journal of Economics*, Autumn 1978, 9, 611-21.
- Sappington, David, "Strategic Firm Behavior Under a Dynamic Regulatory Adjustment Process," *Bell Journal of Economics*, Spring 1980, 11, 360-72.
- \_\_\_\_\_, "Optimal Regulation of Research and Development Under Imperfect Information," *Bell Journal of Economics*, Autumn 1982, 13, 354-68.
- \_\_\_\_\_, "Optimal Regulation of a Multiproduct Monopoly with Unknown Technological Capabilities," *Bell Journal of Economics*, Autumn 1983, 14, 453-63.
- \_\_\_\_\_, and Sibley, David S., "Regulatory Incentive Schemes Using Historic Cost Data," unpublished manuscript, February 1985.
- Sheshinski, Eytan, "Welfare Aspects of the Regulatory Constraint: Note," *American Economic Review*, March 1971, 61, 175-78.
- Sibley, David S. and Bailey, Elizabeth E., "Regulatory Commission Behavior: Myopic Versus Forward Looking," *Economic Inquiry*, April 1978, 41, 249-56.
- Smith, V. Kerry, "The Implications of Regulation for Induced Technological Change," *Bell Journal of Economics and Management Science*, Autumn 1974, 5, 623-32.
- Stigler, George J., "The Theory of Economic Regulation," *Bell Journal of Economics and Management Science*, Spring 1971, 2, 3-21.
- Takayama, Akira, "Behavior of the Firm Under Regulatory Constraint," *American Economic Review*, June 1969, 59, 255-60.
- Vogelsang, Ingo and Finsinger, Jörg, "A Regulatory Adjustment Process for Optimal Pricing by Multiproduct Monopoly Firms," *Bell Journal of Economics*, Spring 1979, 10, 157-71.
- Wellisz, Stanislaw H., "Regulation of Natural Gas Pipeline Companies: An Economic Analysis," *Journal of Political Economy*, February 1963, 81, 30-43.
- Wilson, James Q., *The Politics of Regulation*, New York: Basic Books, 1980.
- Zajac, Edward E., "A Geometric Treatment

of Averch-Johnson's Behavior of the Firm Model," *American Economic Review*, March 1970, 60, 117-25.

\_\_\_\_\_, "Note on 'Goldplating' or 'Rate Base Padding'," *Bell Journal of Economics and Management Science*, Spring 1972, 3,

311-15.

Ziemba, William T., "The Behavior of a Firm Subject to Stochastic Regulatory Review: Comment," *Bell Journal of Economics and Management Science*, Autumn 1974, 5, 710-12.



# When Actions Speak Louder Than Prospects

By GRAHAM LOOMES\*

*Many theories of individual choice under risk and uncertainty are formulated in terms of preferences over prospects, that is, probability distributions of consequences. By contrast, regret theory is formulated in terms of actions, that is, n-tuples of state-contingent consequences. From the viewpoint of prospect-based theories, what appear to be innocuous rephrasings of choice problems are predicted by regret theory to cause people to reverse their choices. This paper follows up earlier results with a new kind of experimental test.*

Consider two pairwise choice problems of the kind illustrated in Figure 1. Columns represent states of the world, likely to occur with the probabilities shown, where  $1 > q > p > 0$ ; rows represent actions, with monetary consequences such that  $a > b > 0$ .

Many theories of choice under uncertainty treat these two problems as equivalent, since in both cases *A* offers *a* with probability *p* and 0 with probability  $1 - p$ , while *B* offers *b* with probability *q* and 0 with probability  $1 - q$ . It is therefore assumed that individuals who prefer *A* in one case must prefer *A* in the other; similarly for *B*. This is true not only for conventional expected utility theory, but for a wide range of other models which are formulated in terms of preferences over prospects (that is, alternatives expressed simply as probability distributions of consequences).

However, there is now a considerable amount of experimental evidence which suggests that individuals are less likely to choose *B* in Problem 0 than in Problem 1. Loomes, forthcoming 1989, presents evidence relating to 14 such pairs of problems drawn from

PROBLEM 0		P	q
	A	a	0
	B	0	b

PROBLEM 1		p	q-p	1-q
	A	a	0	0
	B	b	b	0

FIGURE 1. TWO PAIRWISE PROBLEM FORMATS

experiments involving a total of 542 individuals, all operating on the basis of real monetary incentives. In all 14 cases, *B* was chosen more frequently in Problem 1: the null hypothesis was rejected with at least 99 percent confidence in 7 of the 14 cases, and with at least 95 percent confidence in another 4 cases. A separate study by Chris Starmer and Robert Sugden (1987) provides further evidence of this kind, based on the incentive-linked choices of another 283 individuals; additional problems presented to a subset of 120 individuals show that the pattern extends into the domain of losses.

This evidence presents a serious challenge to prospect-based theories, but it is consistent with, and predicted by, a group of "regret" models such as those proposed by David Bell (1982); Peter Fishburn (1982); and Loomes and Sugden (1982, 1987). In this paper I shall use the most recent of these formulations of regret theory to generate the relevant predictions.

\*Department of Economics and Related Studies, University of York, York YO1 5DD, U.K. The experimental work reported in this paper was funded by the Economic and Social Research Council (U.K.) Award B 00 23 2163. My thanks to Suky Thompson for computer programming, and to David Butler and Norman Spivey for assisting with analysis of the data. Many of the ideas in this paper stem from my collaboration with Robert Sugden. This paper has also benefited considerably from a number of constructive suggestions by John Conlisk.

Consider the slightly more general case where there are two actions  $A$  and  $B$  and several possible states of the world. Let  $p_j$  denote the probability that state  $j$  will occur, and let the respective consequences of  $A$  and  $B$  under the  $j$ th state be denoted  $x_{Aj}$  and  $x_{Bj}$ . The psychological intuition behind regret theory is that if an individual chooses  $A$  (and therefore rejects  $B$ ) and state  $j$  occurs, the overall level of satisfaction he experiences will depend not simply upon  $x_{Aj}$  but also upon how  $x_{Aj}$  compares with  $x_{Bj}$ . If what he gets is worse than what he might have had, it is suggested that the satisfaction associated with  $x_{Aj}$  will be reduced by a decrement of utility due to *regret*; correspondingly, if  $x_{Aj}$  is better than  $x_{Bj}$ , there will be an increment of utility ascribed to *rejoicing*. It is assumed that individuals anticipate all such increments or decrements when making their decisions.

This can be represented most compactly by a function  $\Psi(x_{Aj}, x_{Bj})$ , which measures the *net advantage* of choosing  $A$  and rejecting  $B$  in the event that state  $j$  occurs. If individuals choose so as to maximize expected net advantage, and if we denote strict preference, weak preference, and indifference by  $>$ ,  $\geq$ , and  $\sim$ , respectively, regret theory's decision rule can be expressed as follows:

$$(1) \quad A \succeq B \Leftrightarrow \sum p_j \Psi(x_{Aj}, x_{Bj}) \geq 0.$$

In the special case where  $\Psi(x_{Aj}, x_{Bj}) = U(x_{Aj}) - U(x_{Bj})$ , Expression (1) reduces to the conventional expected utility decision rule. For the three consequences in Figure 1, this would entail  $\Psi(a, 0) = \Psi(a, b) + \Psi(b, 0)$ . However, regret theory suggests that in general this equality will not hold, and that it will often be the case that  $\Psi(a, 0) > \Psi(a, b) + \Psi(b, 0)$ . The psychological interpretation of this inequality (referred to in Loomes and Sugden, 1987, as the *convexity condition*) is that the regret-rejoice effect of a single large difference between consequences is more intense than the sum of the regret-rejoice effects if the difference is broken into two smaller parts.

Now consider an individual who is indifferent between  $A$  and  $B$  in Problem 0.

Applying Expression (1) gives

$$(2) \quad A \sim B \Leftrightarrow p\Psi(a, 0) + q\Psi(0, b) = 0.$$

We can then apply Expression (1) to Problem 1. Invoking the skew-symmetry property, whereby  $\Psi(b, 0) = -\Psi(0, b)$ , and rearranging, gives

$$(3) \quad A \succeq B \Leftrightarrow p[\Psi(a, b) + \Psi(b, 0)] + q\Psi(0, b) \geq 0.$$

If the convexity condition holds,  $[\Psi(a, b) + \Psi(b, 0)] < \Psi(a, 0)$ . Thus indifference in Problem 0 entails  $A < B$  in Problem 1. By a symmetrical argument, an individual who is indifferent between  $A$  and  $B$  in Problem 1 will strictly prefer  $A$  in Problem 0. Thus the regret model generates an alternative hypothesis to prospect-based theories—an alternative which is supported by the evidence cited above.

However, experiments involving particular predetermined sets of values of  $a$ ,  $b$ ,  $p$ , and  $q$  are likely to understate the extent of any regret effect. The model entails that those who prefer  $B$  in Problem 0 will also prefer  $B$  in Problem 1, and that those who prefer  $A$  in Problem 1 will also prefer  $A$  in Problem 0. So even if the convexity condition holds for all members of a sample, only those with a sufficiently weak preference for  $A$  in Problem 0 and for  $B$  in Problem 1 will be observed to switch at any given set of values of  $a$ ,  $b$ ,  $p$ , and  $q$ . The other members of the sample might switch at different levels of parameter values, but experiments to date have not taken a form that could demonstrate this. The experiment described below set out to address that issue. By testing the model in a different way, it aimed to reveal much more about the full extent of any regret effect.

### I. The Experiment

The principle behind the new experiment is as follows. Instead of fixing all parameters and asking participants to make choices, let us fix  $a$ ,  $p$ , and  $q$ , and ask each individual to set the value of  $b$  at whatever level makes

him indifferent between  $A$  and  $B$ . When this is done using a Problem 0 format, let the value which produces indifference be denoted  $b_0^*$ ; when done with a Problem 1 format, let the value be denoted  $b_1^*$ . Prospect-based theories predict  $b_0^* = b_1^*$ , while regret theory predicts  $b_0^* > b_1^*$ . Thus, the distribution of  $b_0^*$  and  $b_1^*$  values across a sample of participants allows a test of regret theory against a broad class of prospect-based models.

By the same token, we can fix  $b$ ,  $p$ , and  $q$  and elicit values of  $a$  which produce indifference:  $a_0^*$  when using a Problem 0 format, or  $a_1^*$  when using the format of Problem 1. Again, prospect-based models predict  $a_0^* = a_1^*$ ; regret theory predicts  $a_0^* < a_1^*$ .

The experiment involved a total of 128 participants, distributed randomly between four subsamples, here labeled  $W$ ,  $X$ ,  $Y$ , and  $Z$ . Each participant was asked to set one  $a_i^*$  value and one  $b_i^*$  with a different combination for each subsample, as follows:

Subsample  $W$ :  $a_0^*$  and  $b_0^*$ .

Subsample  $X$ :  $a_0^*$  and  $b_1^*$ .

Subsample  $Y$ :  $a_1^*$  and  $b_0^*$ .

Subsample  $Z$ :  $a_1^*$  and  $b_1^*$ .

In all cases,  $p$  and  $q$  were set at 0.4 and 0.6, respectively. For questions eliciting values of  $a_i^*$ ,  $b$  was fixed at £10.00; for questions eliciting values of  $b_i^*$ ,  $a$  was fixed at £15.00.

Participants were seated at separate computer terminals and were taken through three practice questions (identical for all subsamples) to give them an opportunity to familiarize themselves with the equipment and the experimental procedure. They were told that after the practice period they would be presented with a sequence of questions. Their decisions would be stored in the computer's memory, and when all questions had been answered, one would be selected at random: their decision in that problem would be replayed on their terminal screen, and their entire payment for taking part in the experiment would depend on how that decision worked out. This arrangement allows responses to any subset of questions to be treated as independent decisions.<sup>1</sup>

<sup>1</sup>Charles Holt (1986) has raised doubts about the reliability of this procedure for eliciting "true" re-

	1	40:41	60:61	100
A	15.00	0.00	0.00	
B	12.00	12.00	0.00	
	40	20	40	

FIGURE 2. INITIAL DISPLAY TO ELICIT  $b_1^*$

Each participant answered twelve questions in all, only two of which were of the type reported in this paper.<sup>2</sup> To illustrate how these questions operated, consider Figure 2 which reproduces the opening display presented to subsamples  $X$  and  $Z$  to elicit  $b_1^*$ . Initially, participants were asked to make a choice by typing in either  $A$  or  $B$ . If  $A$  was chosen, the computer responded by replacing the 12.00 by a larger sum (in this case 13.70) and asked the participant to make a fresh choice; if  $B$  was chosen, a smaller sum (8.60) replaced the 12.00, and participants chose again. After a sequence of, at most, four such choices, the nonzero consequences in  $B$  were replaced in the display by question marks. Underneath, a message reminded the participant of the largest conse-

sponses. He argues that if the independence axiom breaks down, choices or stated values in one question may be distorted by some interaction with other questions in the experiment. This is an important issue with wide implications that requires a larger and more specific empirical examination than the present experiment can provide. However, the present experimental design does take some account of the problem. The question which elicited values of  $b_i^*$  was always positioned second in the sequence. Up to that point, each subsample had seen exactly the same three practice questions and the same first "real" questions, and all participants were equally ignorant of the questions still to come. So even if there were some interaction effect, the logic of Holt's (prospect-based) argument is that this should have the same impact on both  $b_0^*$  and  $b_1^*$ : randomization across subsamples and the fact that at that time each subsample had been given identical information means that systematic differences between  $b_0^*$  and  $b_1^*$  cannot be accounted for by Holt's conjecture. The question eliciting values of  $a_i^*$  always came sixth in the sequence. The third, fourth, and fifth questions *did* differ somewhat between subsamples. However, in fn. 4 in the next section I shall indicate why it is very unlikely that the observed patterns of  $a_0^*$  and  $a_1^*$  could be explained by those differences.

<sup>2</sup>The other questions addressed several different issues: full details are available on request.

quence that had been rejected in favor of  $A$  and the smallest sum that had led to  $B$  being chosen. The participant was then asked to type in the smallest sum within that range which would be just enough to make the person willing to accept  $B$ .

During the practice period it was explained that if a question of this kind were selected to be replayed for real, the amount that would be offered in place of the question marks would be determined by a random draw from a pack of cards, each of which had a different sum of money printed on it.

The rule was that if the sum on the card was equal to or greater than the participant's "indifference value," the nonzero consequence(s) would be set at the full amount on the card, and the individual would be deemed to have chosen  $B$ . Alternatively, if the card offered less than the stated indifference value, the participant would be deemed to have chosen  $A$ .

The reasons why this procedure meant that it was in each participant's interests to state the indifference value as honestly and precisely as possible was explained, using a simplified version of the argument first proposed by Gordon Becker, Morris DeGroot, and Jacob Marschak (1964).<sup>3</sup> It was pointed out that since their final value was constrained to lie within the range determined by their previous sequence of choices, participants should make each of these choices as carefully and accurately as they could.

The last element of the experiment concerns the determination of the state of the world. Prior to the practice period, each

participant had chosen a sealed brown envelope from a pile of 100 such envelopes, each containing a different numbered cloak-room ticket. The numbers set into the top row of hyphens in the Figure 2 display relate to these tickets: an individual playing out action  $A$  in that display would have received £15.00 if the person's envelope turned out to contain a ticket between 1 and 40 inclusive; alternatively, if the pack of cards generated an offer greater than or equal to the stated indifference value, the individual would receive the full amount of the offer if the ticket turned out to be a number between 1 and 60. The numbers at the base of each column showed at a glance how many chances out of 100 there were that a particular state of the world would occur.

Because the experimental design involved several elements, participants were also provided with some brief written notes summarizing the key points; queries were not restricted to the practice period—participants were encouraged to ask for clarification at any time; at every decision point, the computer program gave participants the opportunity to change their minds or correct errors before confirming a decision; and at the end, while each participant watched one of the questions being replayed for real, the experimenter had an opportunity to discuss the experiment with them. The indications were that participants generally understood what they were being asked to do, and gave considered responses.

## II. Results and Discussion

Table 1 shows the means and standard deviations of the distributions of values elicited from each subsample. (Full details of the values given by all 128 participants are reported in the Appendix in Table A1.) Table 1 shows that whenever any two subsamples were presented with identical problems then, as all models would predict, there was no significant difference between the subsample means. By contrast, whenever those same two subsamples were presented with the other question, where they each faced a different problem format, the difference between the two means was always significant,

<sup>3</sup>Edi Karni and Zvi Safra (1987) have shown that under certain assumptions, failure of the independence axiom will result in the Becker-DeGroot-Marschak device eliciting "false" stated values. This possibility, which has yet to be either established or refuted empirically, requires a more specific study. However, even if it is valid, the Karni and Safra argument has no bearing on the central issue in this paper. Their argument is prospect-based, and therefore any under/overstatement of values should affect  $b_0^*$  and  $b_1^*$  equally; and likewise for  $a_0^*$  and  $a_1^*$ . Any differences between  $b_0^*$  and  $b_1^*$ , or between  $a_0^*$  and  $a_1^*$ , cannot be explained by Karni and Safra's (1987) argument.

TABLE 1—MEANS AND STANDARD DEVIATIONS OF ELICITED VALUES

	Subsample			
	W	X	Y	Z
	$a_0^*$	$a_0^*$	$a_1^*$	$a_1^*$
Mean	17.72	17.22	22.24	22.91
Standard Deviation	3.58	5.42	5.34	7.05
	$b_0^*$	$b_1^*$	$b_0^*$	$b_1^*$
Mean	8.57	7.36	8.61	7.19
Standard Deviation	2.26	2.47	2.33	2.13

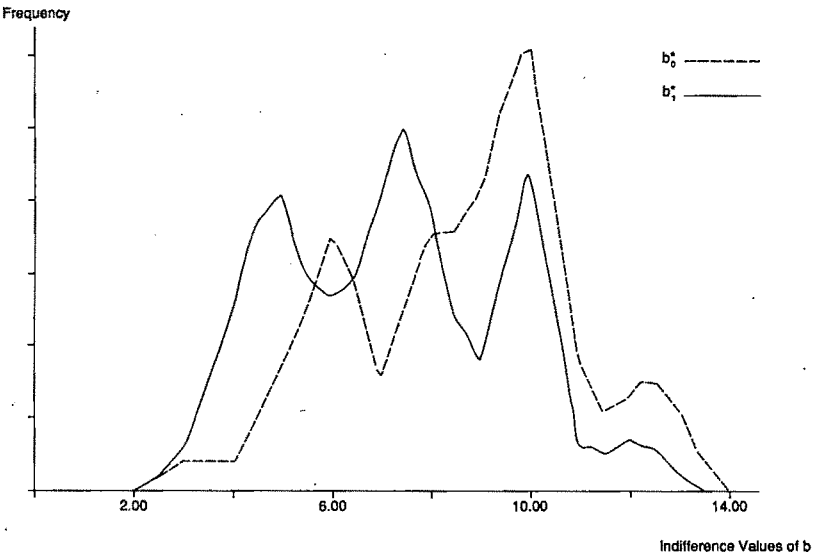
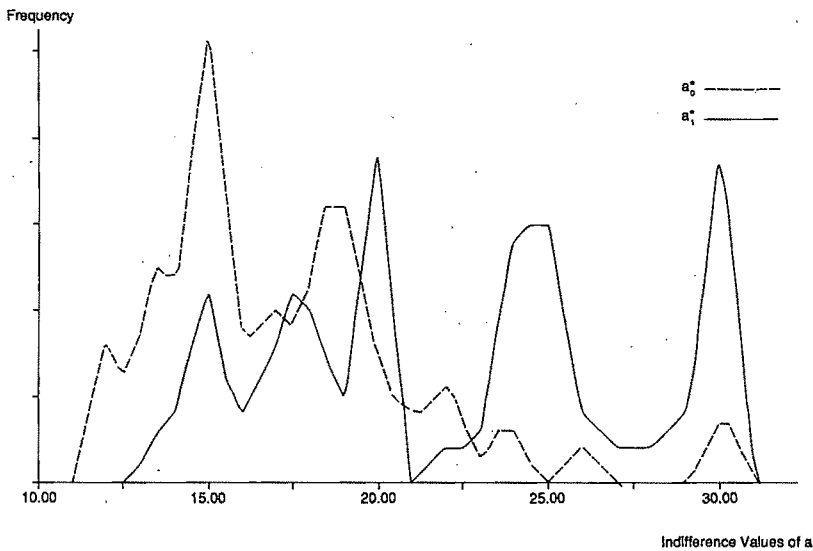


FIGURE 3. SMOOTHED DISTRIBUTIONS OF INDIFFERENCE VALUES

and in the direction predicted by regret theory ( $p < 0.025$  for both comparisons between  $b_0^*$  and  $b_1^*$ ,  $p < 0.005$  for both comparisons between  $a_0^*$  and  $a_1^*$ ).<sup>4</sup>

The extent and significance of observable violations by individual participants is discussed in the Appendix. However, a primary objective of the present study was to reveal more about the extent of any regret effect among those who may not be observed to be violating standard assumptions at any one particular set of values of  $a$ ,  $b$ ,  $p$ , and  $q$ . To obtain an impression of the impact of regret right across whole subsamples, we can pool the observations appropriately and graph the distributions. Because of a tendency for many individuals to round their stated values to the nearest half or whole pound, graphical presentation benefits from smoothing the data somewhat.<sup>5</sup> The results are shown in Figure 3.

Even after smoothing, the distributions remain rather "lumpy." This may be due to sampling variability, or to some other factor(s). Whatever the reason, it does not appear to diminish the regret effect: the predicted shifts in the distributions are neither slight nor confined to some narrow band,

<sup>4</sup>Let us return to the issue discussed in fn. 1. If there was a significant interaction effect between the sixth question and the earlier questions which differed across subsamples, it would have to operate in three particular ways: (i) it would have to cause subsamples  $W$  and  $X$ , which gave significantly different valuations of  $b_0^*$  and  $b_1^*$ , to give the same values of  $a_0^*$ ; (ii) it would have to cause subsamples  $Y$  and  $Z$ , which also gave significantly different valuations of  $b_0^*$  and  $b_1^*$ , to give the same values of  $a_1^*$ ; (iii) at the same time, it would have to cause a significant difference (in the particular direction predicted by regret) between  $a_0^*$  and  $a_1^*$ . Readers who request full details will be able to see how the third, fourth, and fifth questions differed across subsamples, but I can find nothing in those differences which could produce (i), (ii), and (iii) simultaneously as the result of the kind of interaction conjectured by Holt (1986).

<sup>5</sup>The smoothed graphs were produced as follows. All observations were sorted into classes £0.25 wide, with the dividing lines at 0.125, 0.375, 0.625, and 0.875 in every pound. For each observation in a particular class, a weight of  $4/16$  was given to that class; weights of  $3/16$  to the two classes on either side; weights of  $2/16$  to the next two classes, and  $1/16$  to the two classes beyond those.

but persist across almost the entire range of the distributions.

What conclusions should be drawn? This is only one study, and its findings need to be confirmed—or contradicted—by other studies. Nevertheless, the consistency with earlier results gives reason to think that the experimental design reported here may be a useful instrument for exploring other regions of the parameter space and testing other predictions. Meanwhile, it provides further evidence to suggest that decision experiments which fail to control for the juxtaposition of consequences may be neglecting a significant factor, and that the broad class of theories which are framed in terms of prospects rather than actions may all be failing to capture an important element in decision making under uncertainty.

#### APPENDIX

Although the experiment was designed primarily to investigate the extent of the regret effect across whole subsamples, the fact that the probabilities  $p$  and  $q$  were common to both questions allows some examination of observed "violations" by individual participants.

Prospect-based theories assume that for any individual there exists a unique value (denoted simply by  $a^*$ ) which will produce indifference in questions where  $b$  is fixed, irrespective of the problem format; and likewise, a unique value (denoted  $b^*$ ) will produce indifference when  $a$  is fixed. Using conventional notation, we can write a pair of simultaneous equations:

$$(A1a) \quad 0.4U(a^*) + 0.6U(0) \\ = 0.6U(£10.00) + 0.4U(0),$$

$$(A1b) \quad 0.4U(£15.00) + 0.6U(0) \\ = 0.6U(b^*) + 0.4U(0).$$

This entails

$$(A2) \quad a^* \geq £15.00 \Leftrightarrow b^* \leq £10.00.$$

Any individual whose pair of answers fail to satisfy (A2) would be violating conventional theory. Charitably, this might be regarded as "an error."

In regret theory, (A2) holds generally only if the problem format is the same for both questions. That is, regret theory entails

$$(A3a) \quad a_0^* \geq £15.00 \Leftrightarrow b_0^* \leq £10.00,$$

$$(A3b) \quad a_1^* \geq £15.00 \Leftrightarrow b_1^* \leq £10.00.$$

TABLE A1—INDIFFERENCE VALUES OF ALL 128 PARTICIPANTS FOR BOTH QUESTIONS

Individual	Subsample W		Subsample X		Subsample Y		Subsample Z	
	$a_0^*$	$b_0^*$	$a_0^*$	$b_1^*$	$a_1^*$	$b_0^*$	$a_1^*$	$b_1^*$
1	14.67	9.57	17.00	9.46	25.00	10.00	20.00	8.70
2	19.00	11.00	14.00	9.00	20.00	7.50	24.00	6.00
3	19.00	6.00	14.50	6.00	27.00	5.00	23.00	6.00
4	17.50	6.50	17.00	4.00	15.00	6.00	16.00	10.00
5	15.50	10.00	24.00	7.50	17.50	10.00	15.00	10.50
6	19.50	6.25	18.50	5.00	30.00	3.00	22.10	6.10
7	13.34	9.00	14.00	4.00	28.00	11.00	24.00	5.00
8	15.00	10.00	16.00	10.01	30.00	8.60	25.00	10.00
9	13.00	13.00	14.49	3.90	20.00	10.00	30.25	7.49
10	18.50	9.50	40.00	4.99	20.00	10.50	18.60	8.45
11	15.00	10.00	14.00	7.90	17.50	11.00	30.00	3.00
12	22.35	9.85	15.00	6.90	24.50	9.00	18.00	7.00
13	18.05	6.00	16.50	10.20	30.00	6.00	25.00	4.50
14	19.00	5.00	16.00	8.00	29.00	12.20	23.40	7.25
15	20.00	6.50	18.50	5.00	25.00	8.00	50.00	6.00
16	13.60	9.50	15.00	12.50	25.00	6.50	23.90	7.60
17	15.01	10.01	13.00	7.00	25.00	8.75	19.00	9.20
18	19.25	8.05	26.00	7.00	15.01	5.00	30.00	7.50
19	19.00	6.00	12.00	8.00	20.00	10.00	17.50	7.50
20	22.00	7.25	15.00	10.00	33.00	4.00	30.00	4.00
21	17.00	13.00	20.00	7.00	19.90	8.00	26.00	4.50
22	30.00	12.00	15.01	10.01	19.50	8.50	18.50	8.00
23	12.00	9.00	12.00	12.00	18.10	9.00	30.00	6.00
24	19.00	10.50	15.25	6.80	24.00	12.10	25.00	5.00
25	21.00	5.50	21.00	10.00	14.50	9.00	15.01	10.00
26	17.00	8.20	20.00	5.00	15.01	10.01	16.50	11.20
27	13.50	10.00	12.00	7.50	13.50	12.50	17.50	9.00
28	22.00	6.00	15.30	5.00	16.55	9.40	20.00	10.00
29	18.25	5.00	18.00	3.50	26.00	9.00	20.00	6.50
30	18.00	7.50	15.00	10.00	20.00	10.00	29.00	5.00
31	15.00	7.50	13.50	4.25	24.00	8.00	17.00	5.00
32	16.00	10.50	23.41	8.00	24.00	8.00	14.00	8.00

Since regret theory also predicts  $a_0^* < a_1^*$  and  $b_0^* > b_1^*$ , (A3a) and (A3b) allow the possibility of two cases which violate (A2), namely:  $a_0^* < £15.00$  with  $b_1^* < £10.00$ ; and  $a_1^* > £15.00$  with  $b_0^* > £10.00$ . However, regret theory concurs with conventional theory in regarding other violations of (A2) as errors.

Because many participants tended to round their stated values to the nearest pound, it is possible that some observed violations of (A2) might occur simply as a result of rounding error. (For example, if my true values of  $a_1^*$  and  $b_1^*$  are £15.75 and £9.75, respectively, but I round up to £16.00 and £10.00: I then appear to commit a violation.) To reduce the number of such observations, let us define the two "violations" more stringently, as follows

$$V1: a^* \leq £15.00 \text{ and } b^* \leq £10.00$$

$$\text{and } (a^* + b^*) < £24.00,$$

$$V2: a^* \geq £15.00 \text{ and } b^* \geq £10.00$$

$$\text{and } (a^* + b^*) > £26.00.$$

TABLE A2—INDIVIDUAL "VIOLATIONS" BY SUBSAMPLE

Violation	Subsamples			
	W	X	Y	Z
V1	5	10	2	1
V2	5	4	10	3

Prospect-based theories would regard all observations of V1 and V2 in any of the four subsamples as being "errors." However, in just two cases—observations of V1 in subsample X where participants were asked to state  $a_0^*$  and  $b_1^*$ , and observations of V2 in subsample Y where  $a_1^*$  and  $b_0^*$  were being elicited—regret theory would regard such behavior as consistent with the model. See Table A1 for the indifference values of all 128 participants.

Table A2 shows the numbers of participants whose behavior corresponded with either V1 or V2. There were

40 such individuals altogether, and the prospect-based null hypothesis is that only 10 of these should be accounted for by the two cases singled out by regret theory. But what we find is that those two cases actually account for as many observations as the other six cases put together. Although the numbers involved are relatively small, this difference is statistically significant ( $p < 0.01$ ).

## REFERENCES

- Becker, Gordon, M., DeGroot, Morris H. and Marschak, Jacob, "Measuring Utility by a Single Response Sequential Method," *Behavioral Science*, July 1964, 9, 226-32.
- Bell, David E., "Regret in Decision Making Under Uncertainty," *Operations Research*, September 1982, 30, 961-81.
- Fishburn, Peter C., "Nontransitive Measurable Utility," *Journal of Mathematical Psychology*, August 1982, 26, 31-67.
- Holt, Charles A., "Preference Reversals and the Independence Axiom," *American Economic Review*, June 1986, 76, 508-15.
- Karni, Edi and Safra, Zvi, "'Preference Reversal' and the Observability of Preferences by Experimental Methods," *Econometrica*, May 1987, 55, 675-85.
- Loomes, Graham, "Predicted Violations of the Invariance Principle," in P. Fishburn and I. LaValle, eds., *Choice Under Uncertainty*, in *Annals of Operations Research*, special volume, forthcoming 1989, Lucerne: J. C. Balcer.
- \_\_\_\_\_ and Sugden, Robert, "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty," *Economic Journal*, December 1982, 92, 805-24.
- \_\_\_\_\_ and \_\_\_\_\_, "Some Implications of a More General Form of Regret Theory," *Journal of Economic Theory*, April 1987, 41, 270-37.
- Starmer, Chris and Sugden, Robert, "Experimental Evidence of the Impact of Regret on Choice Under Uncertainty," mimeo., University of East Anglia, 1987.



# Limited Contract Enforcement and Strategic Renegotiation

By GUR HUBERMAN AND CHARLES KAHN\*

*This paper presents a strategic theory of contract renegotiation. In this theory, suboptimal contracts are put in place initially to protect one party against undesirable actions by another party and are renegotiated once the danger is past. We develop a model to establish the cases in which simple contracts cannot achieve desirable outcomes, so that only a complicated contract or renegotiation will serve. Unlike most previous accounts of contract renegotiation, this theory does not rely on exogenous uncertainty to motivate renegotiation.*

Contract renegotiation is an agreement to alter an original contract and treat the agreement as the new contract. Contract renegotiation is puzzling: it would seem the renegotiated contract could be anticipated when the original contract is drawn. Therefore, the renegotiated contract could be written into the original contract in the first place, eliminating the need to renegotiate.

Contract renegotiation is even more puzzling when considering contracts which include apparently unreasonable clauses—clauses which are patently suboptimal and which will be renegotiated away rather than carried out.

Bank loans provide an example of contracts which contain seemingly unreasonable clauses. They generally stipulate that assets will be taken over by the bank should the borrower not repay. But banks are usually less efficient as managers of assets than are the borrowers. And in fact, when a borrower ends up short, loans are renegotiated. The threat of takeover is almost never carried out and both parties to the loan realize this. Why then go through the ritual of including such a provision?

Contract renegotiation is often bundled with myopic behavior in general and attrib-

uted to bounded rationality. Given limited computational powers, the contracting parties choose to specify only their most likely actions in the most likely contingencies. If an exceptional circumstance arises, they renegotiate the original contract.<sup>1</sup> This is a poor explanation of loan contracts; in most loan defaults the bank does not foreclose, yet foreclosure is the action specified in loan contracts.

This paper provides a strategic approach to the choice of clauses in contracts in anticipation of their eventual renegotiation. Without disagreeing with the importance of bounded rationality, we suggest that renegotiation can be compatible with complete rationality and play a strategic role in determining the outcome of a relation. We argue that some clauses are included in contracts not because either side expects for them to be carried through, but because the threat of carry through serves a useful strategic purpose. Moreover, in extreme circumstances we show that the best alternative may be to contract explicitly for an undesirable future action, and then renegotiate before the time comes to carry out the action.

We present two examples of strategic renegotiations. In the first, tension between adverse selection and risk sharing is resolved via contract renegotiation. This example can be interpreted as a model of starting up a

\*Graduate School of Business, University of Chicago, and Department of Economics, University of Chicago, Chicago, IL 60637, respectively. We are grateful to Oliver Hart, David Kreps, Robert C. Marshall, Eric Maskin, and Menachem Mautner for helpful comments and to the National Science Foundation for grant no. SES-8511137, and the University of Chicago's Center for Research in Security Prices for financial support.

<sup>1</sup>This is the approach to contract renegotiation taken, for example, in Milton Harris and Bengt Holmstrom, 1987; Ronald Dye, 1985; and Steven Shavell, 1984.

firm (or management's leveraged buy-out of an existing firm) and a subsequent equity sale to the public.

The second example contains no exogenous uncertainty. It is leaner but less realistic than the first. It is included for two reasons. First, by eliminating uncertainty we are able to make a sharp distinction between our strategic account of renegotiation and other accounts which rely on surprises. Second, the simplicity of the example allows an exhaustive study of the possible outcomes.

Our account of strategic renegotiation requires two ingredients: a limit to contract enforceability and a cost to contract complexity. The limit to enforceability is modeled as the enforcement agency's (the court's) inability to observe some of the parameters on which the contracting parties would like to contract. These parameters are called unverifiable. Their presence restricts the set of feasible contracts.

In the presence of unverifiable actions, complicated procedures may be necessary to elicit optimal actions. Contractual renegotiation, however, is a partial substitute for contractual complexity. We define a restricted set of feasible contracts corresponding to a limitation on the number of contingencies permitted in a contract.

As the values of the unverifiable parameters are realized, the contracting parties' preferences over the feasible set of contracts change. Moreover, both the existing contract and the anticipation of future ones affect the parties' incentives. The result is a natural role for renegotiation.

Related work is reviewed in Section I. The example with exogenous uncertainty is presented in Section II. The example with no uncertainty is presented and analyzed beginning in Section III. Section IV provides investigation of the model, and Section V contains the concluding remarks.

### I. Literature Review

The relation between the parties in our model lasts two periods. Parameters realized in the first period are not observable to third-party enforcers. This basic structure appears in a variety of papers which are

concerned with opportunistic behavior in the second period and its effect on parties' first-order actions and their incentives to establish binding relationships.

Benjamin Klein and Keith Leffler (1981) consider a relation between a seller and buyers of a good whose quality is not verifiable by a court. Therefore, the seller cannot bundle the good with a quality warranty. If the market expects the good to be of high quality but the seller offers a low-quality good to the unsuspecting buyers, he can reap a short-run profit. The quality can be guaranteed if the sale is repeated and the seller develops a reputation for high quality. The reputation is developed by an initial investment in an asset dedicated to the good's production. Philip Dybvig and Chester Spatt (1985) and Carl Shapiro (1982) make similar arguments.

Oliver Williamson (1983) suggests that two parties to a relation exchange hostages as a guarantee against opportunistic behavior. In Williamson's finite horizon model, the threat of the loss of a hostage suffices to eliminate opportunistic behavior. Asset specificity (human or physical capital) and asset location are examples of hostages. The value of such investments depends on the continuation of the relation. Vincent Crawford (1982) develops similar arguments, and Jean Tirole (1986) extends this framework in a stochastic model.

Carliss Baldwin (1983) considers a labor union's ability to extract concessions from an employer once the employer irreversibly invests in a production technology. She argues that, anticipating the union's behavior, the firm makes an inefficient choice of capital. Baldwin notes that the inefficiency disappears if the union buys out the firm.

It is tempting to generalize this observation to argue that any enlargement of an organization reduces opportunistic behavior, and if the enlargement is sweeping enough, it eliminates all opportunistic behavior. Sanford Grossman and Oliver Hart (1986) refute this last tempting generalization. Their model has two periods, verifiable and unverifiable actions, and trade in the right to choose second-period actions. Integration of a firm is defined as unified ownership of second-period rights. They show that in-

tegration may or may not be desirable according to the problem's data.

Discussions of contracts in the law-and-economics literature (for example, Shavell, 1980) also use a two-period structure with (implicitly) unverifiable first-period actions (so-called "reliance actions"). This literature focuses on comparing the effects of various damage rules as remedies for breach of contract.

In the law-and-economics literature, authors have observed that legal provisions may only serve as starting points for private renegotiations. Using the example of divorce settlements, Robert Mnookin and Lewis Kornhauser (1979) stress that individuals bargain away from the initially imposed conditions of the legal system. Likewise, William Rogerson (1984), and Shavell (1984) emphasize that parties to a contract can renegotiate terms which are inefficient *ex post*. Shavell concludes that presence of renegotiation makes little difference in the relative efficiency *ex post* of various damage rules.

Our approach is quite different. In our framework there is sufficient flexibility that *ex post* actions are always efficient. We focus instead on the effects that foreknowledge of renegotiation has on initial establishment of contracts, and thus on the efficiency or inefficiency of initial actions. We show that this foreknowledge can be used to create contracts to achieve otherwise unachievable outcomes.<sup>2</sup>

Hart and John Moore (1985) consider an exchange agreement in a model where both the buyer's reservation price and the seller's production cost are uncertain *ex ante* and unverifiable *ex post*. Hart and Moore study the contract written *ex ante* and its *ex post* renegotiation. They conclude that with full-contracting powers in equilibrium no renegotiation is ever required.

<sup>2</sup>Thus, for the most part, the Rogerson-Shavell papers are concerned with issues which are not related to those of interest in this paper. But in some respects the framework Rogerson examines can be regarded as a special case of ours, and one of his results is in apparent contradiction to ours. We will discuss the relation and the reason for the apparent discrepancies at the end of Section IV, Part D, below.

The present paper reaches the opposite conclusion by restricting the set of feasible contracts. We then reconstruct examples in which contract renegotiation is the only route available to the parties to reach the Pareto-optimal play.

## II. An Example with Uncertainty

The example presented here deals with changes in corporate ownership. The example is an adverse-selection model, which can be interpreted as an account of a leveraged buy-out and subsequent resale of the equity to the public. Another application of the example is to a starting enterprise. In it the entrepreneur initially holds a large portion of the equity, but eventually takes the firm public.

The example is related to the work of Hayne Leland and David Pyle (1977) who study the problem of a risk-averse entrepreneur who owns a risky project and wishes to share the risk by selling some of the equity. The entrepreneur has private information pertaining to the project's value, but this information never becomes available to the investor, let alone to the authority that would enforce a contract. We modify the model of Leland and Pyle by assuming that the entrepreneur's private information becomes available to the investor (but not to any third party) after the initial contract is signed. At this point the initial contract is renegotiated.

Consider a risk-neutral investor facing a risk-averse entrepreneur who has utility function  $U$  and a type unknown to the investor,  $a$ . The entrepreneur of type  $a$  has an idea for a project whose ultimate value is a verifiable random variable  $x$  with probability density function  $f(x|a)$ . The support of the random variable  $x$  does not depend on  $a$ . To the investor, all projects look alike, but every entrepreneur  $a$  has some information which leads him to believe that the expected value of his project is  $a$ . As  $a$  increases, the distribution  $f$  improves in the sense of first-order stochastic dominance. Each project requires an investment of \$ $i$ .

Events unfold as follows. A contract is written between the entrepreneur and the

investor and the investment is made. Then  $a$  becomes known to the investor. It is, however, not verifiable. At this point the contract can be renegotiated. Finally,  $x$  is realized and revenues are split according to the final contract.

The socially desired outcome is that a project is taken if and only if  $a > i$ , and that the investor, being risk neutral, bears all the risk. The contract cannot be made contingent on  $a$ , because  $a$  is not verifiable. The socially desired outcome cannot be achieved with any contract that is left intact after  $a$  is known to the investor; if the contract guarantees a sufficiently high fixed payment to the entrepreneur, all projects will be pursued, including those whose  $a < i$ . (If the contract promises too low a payment, no project will be pursued.) If the contract makes the payoff to the entrepreneur contingent on the outcome  $x$ , risk will be shared suboptimally.

The socially desired outcome will be achieved as follows. Initially the contract specifies that the investor receive  $h(x)$  and the entrepreneur receive  $x - h(x)$ . Once  $a$  is observed, the initial contract is renegotiated to a contract that gives the entrepreneur a fixed amount  $A$ , and the investor becomes the residual claimant. Suppose the Nash-bargaining solution is the outcome of the contract renegotiation. Then  $A$  solves

$$(1a) \quad \text{maximize} [U(A) - E\{U(x - h(x))|a\}] \\ \times [E\{(x - A)|a\} - E\{h(x)|a\}].$$

In general,  $A$  depends on  $a$  and on the contract  $h$ ,  $A = A(a; h)$ . Indeed, we show below that if the initially specified payment to the entrepreneur  $x - h(x)$  is increasing in  $x$ , then  $A(a; h)$  is an increasing function of  $a$ . To achieve the socially desired outcome, choose  $h$  so that  $U(A(a; h)) > U(0)$  if and only if  $a > i$ .

**LEMMA:** Suppose  $x - h(x)$  increases with  $x$ . Then  $A(a; h)$  increases with  $a$ .

**PROOF:**

The  $A$  which solves (1a) satisfies

$$(1b) \quad U'(A)[E\{(x - A)|a\} - E\{h(x)|a\}] - [U(A) - E\{U(x - h(x))|a\}] = 0.$$

Differentiate (1b) with respect to  $a$  to obtain

$$(1c) \quad U'(A)(dE\{(x - h(x))|a\}/da) + \{dE\{U(x - h(x))|a\}/da\} + \{dA/da\}[U''(A)[E\{(x - A)|a\} - E\{h(x)|a\}] - 2U'(A)] = 0.$$

Monotonicity of  $x - h(x)$  implies that the first two terms on the left-hand side of (1c) are positive. Also, the derivative  $dA/da$  is multiplied by a negative quantity. (Recall that  $[E\{(x - A)|a\} - E\{h(x)|a\}]$  is positive, being the improvement in the investor's welfare due to the bargain.) Therefore,  $dA/da > 0$ .

When a management team takes a publicly held firm private, it borrows the necessary funds and retains an equity position in the firm. Frequently, the management's position is riskier than straight equity because the lenders are promised a fraction of the (uncertain) profits, for example, by having the debt convertible into equity. At times leveraged buy-outs entail issuing warrants to the management. The value of a manager's human capital is already highly correlated with the fortunes of the firm he manages. Why impose so much additional risk on him? Our example suggests that the burden of risk is temporarily put on the management in order to sort the management teams which have good projects from those which do not have such projects, but might claim they have them.

Sorting cum risk bearing need not last forever. Once the public learns that the managers are not charlatans, it is time to share the risks better; the management sells its equity back to the public.

It is possible to derive a similar result by replacing adverse selection with moral hazard. Suppose all entrepreneurs are of equal ability, so that there is no room for adverse selection. After the initial contract is made and investment takes place, the entrepreneur takes an action which is observed by the investor (but not by any third party). Then the parties renegotiate the initial contract. The initial contract is necessary because the entrepreneur's action is costly to him but beneficial to the enterprise. The risk of the joint enterprise is shared better under the renegotiated contract.

In either formulation the idea is similar: an incentive problem is solved by a temporary allocation of risk to a risk-averse individual. Once the incentive problem is moot, then the associated contract can be renegotiated to reallocate the risk.

### III. An Example with No Uncertainty

In the remainder of this paper, we investigate strategic renegotiation in a deterministic model. We do so for two reasons. First, by eliminating uncertainty we can demonstrate that strategic renegotiation is a phenomenon completely distinct from renegotiation due to surprises. Second, we can derive a problem sufficiently simple to allow an exhaustive analysis of when renegotiation will occur.

The main features of the deterministic model are as follows. Two players, *I* and *II*, have a relation in which *I* chooses between two actions, *a* and *b*, and *II* chooses between two actions, *x* and *y*. The payoffs to *I* and *II* depend on the actions chosen. Both actions are observed by both players. Player *I*'s choice must be made first and it—unlike player *II*'s choice—is not observable to any third party. A contract is a promise of a side payment contingent on an action. As only *II*'s actions are observable to third parties, only they can be enforced in a contract. A contract can be drawn both before and after *I* moves. Moreover, a contract written before *I* moves can be renegotiated after *I* moves. As long as the players do not agree to change an existing contract, the existing contract binds both of them.

The example illustrates the strategic role of contracts in general and of contract renegotiation in particular. A contract written before *I* moves is the disagreement point (in the language of Nash-bargaining models) in the negotiations which follow *I*'s move. Therefore, even when the players agree to alter the original contract, the split of the surplus accruing to the players from writing the second contract depends on the original contract. Although the original contract may be renegotiated, its role in the renegotiations affects eventual payoffs and thereby affects *I*'s choice of action. Thus a judicious construction of the original contract can lead the parties to the socially desired outcome. En route to that outcome the original contract may be renegotiated.

*The Framework.* Two risk-neutral agents are in an economic relation in which agents choose actions sequentially. These actions affect both agents' utilities (payoffs). All actions are observable by the two active agents. A good—"money"—acts as a transferable utility between the two agents.

In addition to the two active agents a third, passive, agent—the "court"—has limited power to enforce contracts made by the other two agents. The court can observe some actions but not others. We will call the former actions "verifiable" and the latter "unverifiable." Transfers of money are observable by the court. Under this assumption, no loss of generality is entailed in assuming that the actual transfers occur only in the last period.<sup>3</sup>

In summary, the technology of a game is specified by a game tree for a sequential and full-information game between two individuals.<sup>4</sup> In addition, all nodes are specified as verifiable or unverifiable. The simplest game

<sup>3</sup>Although we assume that transfers at the terminal node are verifiable, we assume that other aspects of terminal payoffs are unverifiable. Otherwise, by examining terminal payoffs, the court might be able to deduce directly which unverifiable actions had previously taken place.

<sup>4</sup>Although we confine ourselves to situations in which actions occur sequentially, the model's main ideas are applicable to cases where the agents make simultaneous moves. (See Gur Huberman and Charles Kahn, 1985, for examples with simultaneous moves.)

TABLE 1—THE GAME'S BASIC DATA

	II		
	x		y
	a	e, f	g, h
I	b	p, q	r, s

of interest has two unverifiable nodes followed by two verifiable ones. It is depicted in Table 1 and is studied in the rest of the paper.

Player *I* plays first. Player *II* moves after observing *I*'s move. Player *II*'s move is verifiable by a court; player *I*'s move is not. The pairs (*e*, *f*), (*g*, *h*), etc., represent the players' payoffs.

Without loss of generality assume that the pair (*a*, *y*) is the Pareto-optimal move. In other words, *g* + *h* is greater than the sum of the payoffs in any other cell of the matrix.

On the game's technological structure we superimpose the contracting and renegotiating structure. We do so by expanding the game tree: before any physical action of the basic game the players enter a "negotiating round" in which they exchange, and sometimes agree to, contract offers. In this expanded game, the payoff is whatever physical payoff results from the sequence of physical actions of the underlying technological game, plus a transfer equal to that specified in the "active contract"—that is, in the last contract to be proposed and accepted.

The contract that emerges from the negotiations—if one emerges—depends on the protocol the players use to negotiate. The court can observe only the negotiations' outcome (i.e., the contract, if there is one) but not the players' behavior (for example, offers they make) during the negotiations.

A contract specifies the payoff to be made by player *II* to player *I* (it may be positive or negative) as a function of the sequence of verifiable actions made over the course of the game.<sup>5</sup>

<sup>5</sup>We have therefore ruled out more complicated contracts in which individuals' payoffs depend not only on the actions of the players but also on their announce-

TABLE 2—THE GAME'S PAYOFFS IF THE CONTRACT (*X*, *Y*) HOLDS

	II		
	x		y
	a	e + <i>X</i> , f - <i>X</i>	g + <i>Y</i> , h - <i>Y</i>
I	b	p + <i>X</i> , q - <i>X</i>	r + <i>Y</i> , s - <i>Y</i>

In the simple game depicted in Table 1, a contract consists of a pair (*X*, *Y*), interpreted as the following commitment: if *II* plays *x*, he pays *X* to player *I*; if *II* plays *y*, he pays *Y* to player *I*. The side payments *X* and *Y* can be negative. The payoffs of the original game modified by the contract (*X*, *Y*) are in Table 2. In the analysis which follows, a central role is played by the difference *D* = *Y* - *X*.

The simplest contract is for *specific performance*. Such a contract orders player *II* to take a certain action with no recourse. The set of contracts we consider is a broader set. It can be thought of as mandating one of the verifiable actions and imposing *liquidated damages* on player *II* if the other action is chosen. If the damage levels are so high as to ensure player *II* only performs the mandated action, the liquidated damage contract reduces to a specific-performance contract. Note that if both players' actions were verifiable, specific-performance contracts would always suffice to guarantee the optimum.

A contract of particular interest is the "no-contract" contract, that is, the contract assumed operative if no contract is accepted in any particular round. This is of course the contract which specifies no transfer payment no matter what sequence of actions is taken in the game.

Contracts can be renegotiated. The court enforces only the latest of the contracts which

ments (which can in effect be announcements of the unobserved actions). Such contracts can be ruled out on the basis of complexity. For example, if we use the number of clauses in a contract as a measure of complexity, contracts which depend on announcements will require more contingencies than those contracts we consider here.

have been accepted by both individuals. At any stage of the game, the individual empowered to propose contracts can propose a new one; if accepted, that contract supersedes any previous contract. It is assumed impossible to write a contract prohibiting renegotiation.

Subgames can be naturally defined within this structure. In any contract proposed after the game begins, there is no need to specify the payoffs for bygone actions since they cannot be changed. It is thus straightforward to establish that subgame outcomes depend solely on the current contract—that is the one which will be the active contract if no new contract is accepted in the rest of the game—and on the sequence of physical actions taken thus far.

The outcome of contract negotiations and the split of the surplus depend on the negotiations' protocol, that is, the order at which the players make offers and their incentives to conclude the exchange of offers rather than go on indefinitely. The simplest form of negotiation consists of an offer of a contract by one player and an acceptance or rejection of the contract by the other player. These two protocols are summarized as (I.A) and (II.A):

(I.A) In any negotiation round, player *I* can offer one contract and player *II* can accept or reject it.

(II.A) In any negotiation round, player *II* can offer one contract and player *I* can accept or reject it.

Protocols (I.A) and (II.A) seem artificial in the severe restrictions they place on the players' ability to negotiate. The approach of Ariel Rubinstein (1982) provides a more general structure, allowing individuals to exchange offers and counteroffers. One individual makes an offer and the other accepts or rejects it and makes a counteroffer. The bargaining outcome is required to be the perfect equilibrium of the bargaining game.

This approach requires an additional assumption which makes bargaining a costly process, so that the parties have an incentive to settle rather than to continue to haggle. Kenneth Binmore, Rubinstein, and Asher Wolinsky (1986) propose the following interpretation of bargaining costs: the cost of

bargaining arises from the possible termination of the bargaining by an outside power prior to the bargaining's conclusion. Given that negotiations are occurring at time  $t$ , the probability of outside termination before time  $T$  hence is  $1 - \exp\{-L(T-t)\}$ . If the negotiations are terminated by an outside power, the existing contract is binding and the players have no other chance to renegotiate it.

Binmore et al. (1986) demonstrate that as the inter-offer time converges to zero, the outcome of the bargaining game is a split of the surplus in proportions determined by the two players' risk aversion. When the individuals are identical, the split is even. Protocols (I.A) and (II.A) can thus be reinterpreted as extreme versions of this model in which parameters are such that one player appropriates all the gains from cooperation. Other models lie between them in that each party appropriates some of the surplus.

In the next section we study the game assuming that the surplus in renegotiation is split between the parties in proportions  $(c, 1-c)$ .

#### IV. Investigation of the Model

##### A. Outcomes with No First-Period Contract

Suppose no commitment was made in the first period and player *I* plays  $a$ . Evaluate the subgame which ensues. Since the second round's actions are verifiable, the contracting outcome maximizes the players' joint income, given player *I*'s previous choice.

Consider first the case  $h > f$ . Then player *II*'s own choice is also the Pareto optimum. Therefore no acceptable contract will alter the outcome: *I* will not accept any contract which pays him less than  $g$ , and *II* will not offer any contract which assures him less than  $h$ . So the only reachable contracts are "null" contracts—that is, they offer side payments of zero on the condition that player *II* do what he would have done in the absence of the contract.

The players negotiate a contract if  $h < f$ . Recall that failure to agree prior to the outside termination leaves the players without a contract. If this happens after player *I* has

played  $a$ , then  $II$  plays  $x$  and the players receive  $e$  and  $f$ , respectively. Thus the joint gains from the contract are  $(g + h) - (e + f)$ , which are split in proportions  $(c, 1 - c)$ .

In summary, if  $h > f$ , player  $II$  acts unilaterally and the players receive  $(g, h)$ ; if  $h < f$ , the players negotiate an agreement and receive  $(e + c(g + h - e - f), f + (1 - c)(g + h - e - f))$ . Define  $a^*$  as the value to player  $I$  from playing  $a$  when no contract is binding. Then

$$(2a) \quad a^* = g \quad \text{if } h > f,$$

$$(2b) \quad a^* = e + c(g + h - e - f) \quad \text{if } h < f.$$

The definition of  $a^*$  for  $h = f$  is arbitrary; we can make player  $I$ 's payoff (2a) or (2b) by suitably specifying player  $II$ 's response.

The consequences of  $I$  playing  $b$  in the absence of a binding contract are analogous. Namely,

$$b^* = \begin{cases} r & \text{if } r + s > p + q \text{ and } s > q \\ p + c(r + s - p - q) & \text{if } r + s > p + q \text{ and } s < q \\ p & \text{if } r + s < p + q \text{ and } s < q \\ r + c(p + q - r - s) & \text{if } r + s < p + q \text{ and } s > q \end{cases}$$

More succinctly, express  $b^*$  as

$$(3c) \quad b^* = p + c \max\{0, r + s - (p + q)\} \quad \text{if } q > s,$$

$$(3d) \quad b^* = r + c \max\{0, p + q - (r + s)\} \quad \text{if } q < s.$$

Again, for  $q = s$ ,  $b^*$  depends on which of the two responses is attributed to player  $II$ .

In the absence of a first-period contract  $I$  is willing to play  $a$  if and only if  $a^* \geq b^*$ . If this holds and  $a^* = g$ , the Pareto-optimal play is achieved without any contract. If, however,  $a^* \geq b^*$  but  $a^*$  does not equal  $g$ , the players will negotiate a contract after  $I$  plays  $a$ .

Let  $v^*$  be the payoff to  $I$  of the game with no first-period contract. Then

$$v^* = \max\{a^*, b^*\}.$$

If  $v^* = c^*$ , the efficient outcome can be achieved without signing a first-period contract. There are two subcases: If  $v^* = a^*$  and  $a^* = g$ , the efficient outcome can be achieved without signing a second-period contract. In this case the only contracts that can be offered and accepted in period 1 are null contracts.

In the second subcase  $v^* = a^*$  and  $a^* = e + c(g + h - e - f)$ . In this subcase, once player  $I$  has played  $a$ , player  $II$  prefers to play  $x$ . In the second period, the two players will write a contract to induce the Pareto-optimal play.

### B. The Role of First-Period Contracts

Suppose the contract  $(X, X + D)$  is in place before player  $I$  makes his move. The modified game's data are summarized in Table 2. The following analysis of the modified game is analogous to the analysis above.<sup>6</sup>

Denote by  $A^*(X, D)$  and  $B^*(X, D)$  the payoff to player  $I$  if he plays  $a$  and  $b$ , respectively. Let  $A^*(D) = A^*(X, D) - X$  and  $B^*(D) = B^*(X, D) - X$ . Then the analogues to (2a)–(3d) are

$$(4a) \quad A^*(D) = g + D \quad \text{if } h - f > D,$$

$$(4b) \quad A^*(D) = e + c(g + h - e - f)$$

$$\text{if } h - f < D,$$

and

$$(5a) \quad B^*(D) = p + c \max\{0, r + s - (p + q)\} \quad \text{if } s - q < D,$$

$$(5b) \quad B^*(D) = r + c \max\{0, p + q - (r + s)\} + D \quad \text{if } s - q > D.$$

<sup>6</sup>The value of  $D$  is set to accommodate the possibility of contract renegotiation and to induce the first player to make the socially preferred move. The value of  $X$ , however, is determined by the players' relative strengths at the outset of negotiations. This relative strength is a function of the prenegotiation threat point and of the first round negotiation protocol.



Equations (4a)–(5b) reduce to (2a, b)–(3c, d), if  $D = X = 0$ .

### C. *The Role of Contracts in Achieving the Optimum*

Having investigated the consequences of having no first-period contract, and the consequences of having various possible first-period contracts, we are now able to determine for which values of the model's parameters the optimum is achievable by unrenegotiated contracts and for which values of the parameters it is only achievable by renegotiated contracts. Among the contracts we will also be able to distinguish between those that are specific-performance contracts and those for which liquidated damages must be specified. Recall that specific-performance contracts simply specify a single outcome while liquidated damage contracts specify a penalty attached to one action relative to the other. As we will see below, it will not always be possible to implement the optimum with specific-performance contracts.

The optimum is achievable in this model if there exists a  $D$  such that  $A^*(D) \geq B^*(D)$ . If the optimum is achievable with  $D$  arbitrarily large, then a specific-performance contract specifying  $x$  will achieve the optimum. If the optimum is achievable with  $-D$  arbitrarily large, then a specific-performance contract specifying  $y$  will achieve the optimum. If in the optimum  $A^*$  is defined by (4a) no renegotiation is involved. If it is defined by (4b), second-period renegotiation occurs. (If the optimum is achievable with  $D = 0$ , no first-period contract is required.)

From equations (4a, b)–(5a, b), we conclude that there are four cases in which the players can reach the Pareto-optimal outcome:

Case 1: A specific-performance contract to play  $y$  is written at the outset and is not renegotiated. Since this requires  $-D$  arbitrarily large, it occurs if and only if (4a) and (5b) are the relevant equations and  $A^* \geq B^*$ , that is,

$$(5) \quad g \geq r + c \max\{0, p + q - (r + s)\}.$$

Case 2: A specific-performance contract to play  $x$  is written at the outset but is renegotiated to play  $y$  after  $I$  plays  $a$ . Since this requires  $D$  arbitrarily large, it occurs if and only if (4b) and (5a) are the relevant equations and  $A^* \geq B^*$ , that is,

$$(6) \quad e + c(g + h - e - f) \\ \geq p + c \max\{0, r + s - p - q\}.$$

Case 3: A liquidated damage contract is written at the outset which induces  $II$  to play  $y$  if  $I$  plays  $a$ . This contract will not be renegotiated and will lead to the optimal play if and only if (4a) and (5a) are the relevant equations and  $A^* \geq B^*$ . A value of  $D$  exists which satisfies these requirements if and only if

$$(7a) \quad h - f \geq s - q,$$

and

$$(7b) \quad h - f \geq p + c \max\{0, r + s - p - q\} - g.$$

Case 4: A contract which, if adhered to, induces  $II$  to play  $x$  if  $I$  plays  $a$  and to play  $y$  if  $I$  plays  $b$ . This contract is renegotiated and, taking the renegotiation into account, player  $I$  is induced to play  $a$ . After  $I$  moves the contract is renegotiated so  $II$  plays  $y$ . This sequence takes place if and only if (4b) and (5b) are the relevant equations and  $A^* \geq B^*$ . It is feasible if and only if

$$(8a) \quad h - f \leq s - q,$$

and

$$(8b) \quad h - f \leq e + c(g + h - e - f) - r - c \max\{0, p + q - r - s\}.$$

Conditions (8a, b) entail the most complex considerations. To attain the Pareto-optimal play  $(a, y)$ , the parties first write a contract which induces  $I$  to play  $a$ . If he plays  $a$ , it is best for  $II$  to play  $x$  under the terms of the original contract. If he plays  $b$ , it is best for

*II* to play  $y$ . This may seem paradoxical: the original contract is drawn so that if *I* plays cooperatively, *II* prefers to take the non-Pareto optimal action  $x$ , but if *I* takes the non-Pareto optimal action  $b$ , *II* prefers to take the cooperative action  $y$ . The paradox is resolved once we recall the role of contract renegotiation which follows *I*'s move: taking it into account he prefers to play  $a$ .<sup>7</sup>

*Comment.* If player *I* chooses  $b$  we can describe him as "misbehaving"; if he chooses  $a$ , we can describe him as "behaving himself." Cases 1–4 are useful for understanding the role played by contracts in inducing a party to behave himself when third parties cannot monitor him directly.

If the individual would behave himself, provided all other agents behave themselves (corresponding to condition (5)), then the solution is quite simple: commit the others to the correct action. Thus one role for a contract is to secure commitment on observable aspects in a hope that proper behavior will follow on the unobservable aspects.

If the establishment of a commitment on the part of player *II* will not induce proper behavior by player *I*, then perhaps the establishment of a threat will. In this circumstance the contract's role is to skew the payoffs for the second player between retaliation and nonretaliation, thereby making the threat credible. Conditions (7a, b) describe circumstances when this strategy is feasible. A special subcase occurs when  $D = 0$ . Then in the absence of a contract it is in the second player's interest to retaliate if the first misbehaves.

The two strategies above entail no contract renegotiation. If neither of them works, it may still be possible to elicit cooperation by anticipating the contract's renegotiation. If the first player could be induced to behave, provided the second player played the "wrong" response, then cooperative play can be achieved by initially signing a contract

committing the second player to the inefficient play  $x$ , and then renegotiating it once cooperative play from the first player has been achieved. This scheme is effective if (6) holds.

Default provisions in loans work in precisely this way. The threat of taking over the borrower's assets is useful in inducing types of behavior that the bank desires (for example, prudent investment policies), but which it would be difficult for a court to verify. If default does occur, the inferior clause, having served its purpose, can be renegotiated away. This application is investigated in Huberman and Kahn (1988).

#### D. On the Necessity of Renegotiation

We expect to observe renegotiation whenever the outcomes that can be achieved with renegotiation are superior to the outcomes that can be achieved with unrenegotiated contracts. Often, however the same outcome can be achieved either with a renegotiated or an unrenegotiated contract. Since we wish to make predictions as to when renegotiation will be observed, we need to take a stand as to which outcome will be chosen in the case of such ties. We make the natural assumption that renegotiation has a small but positive cost, so that in the case of ties parties will adopt the unrenegotiated contract:

**ASSUMPTION:** *Renegotiation occurs when (and only when) the optimal outcome that can be achieved with renegotiation is not achievable in an unrenegotiated contract.*

**THEOREM 1:** *Renegotiation occurs if and only if (6) holds and (5) and (7a) are violated.*

**PROOF:**

If the optimum cannot be achieved, the outcome of the game will be  $b$  plus whichever of the pair  $(x, y)$  is socially more desirable. This outcome can be achieved without a first-period contract. If the optimum can be achieved by renegotiating a liquidated damage contract (8a, b), it can also be achieved through an unrenegotiated specific-performance contract (5). Thus renegotiation will be observed if and only if (6) holds but

<sup>7</sup>In more general models described below, this fourth case is an independent possibility. In this two-by-two version, the fourth case need never be implemented. Whenever it succeeds an unrenegotiated specific-performance contract will also work.

both (5) and (7a, b) are violated. However, (6) and  $g + h > e + f$  imply (7b); thus it must be (7a) that is violated.

**THEOREM 2:** *As long as  $c < 1$ , there exist values for the payoffs in the game such that renegotiation is necessary. If  $c = 1$  (corresponding to protocol 1.A) anytime the optimum can be achieved with renegotiation, it can also be achieved with an unrenegotiated contract.*

**PROOF:**

If  $c = 1$  the three conditions are inconsistent. If  $c$  is not equal to 1, the following procedure generates a set of payoffs which will require a renegotiated contract: For  $e, f, g, h, r, s$ , pick any values such that  $g + h > e + f$ ;  $g + h > r + s$  and  $r > g$ . Then choose  $q$  sufficiently low that  $q < s - h + f$ . Finally, choose a very low value for  $p$ . By this we mean  $p$  must be sufficiently low that

$$p + q < r + s,$$

$$p < e,$$

$$(1 - c)p < c(q - f) + (1 - c)e.$$

As long as  $c$  is not equal to 1, it is possible to pick up an 8-tuple satisfying these requirements. This 8-tuple satisfies the conditions of Theorem 2.

This procedure also generates natural sufficient conditions for the occurrence of renegotiation:

**THEOREM 3:** *In a matrix in which  $(a, y)$  is the Pareto-optimal move, renegotiation will occur for sufficiently low values of  $p$  if  $r > g$  and  $q < s - h + f$ .*

In other words, substantively, three considerations are sufficient to induce renegotiation. First, suppose that if player II were committed to play the socially optimal action then player I would prefer to misbehave. Second, suppose that the marginal benefit to player II of picking the socially optimal action is greater if player I misbehaves. Finally, suppose that there are enormous costs to player I of misbehaving if

player I commits to the inefficient response. Then renegotiation will occur.

These three considerations can be understood in the following way. The first condition makes an unrenegotiated specific-performance contract impossible, since binding player II leaves player I free to misbehave. The second consideration makes an unrenegotiated liquidated damages contract impossible, since the power of such contracts arises if player II's payoffs make retaliation desirable *ex post* only when player I has misbehaved. Finally, the third consideration makes renegotiation successful, for it guarantees that should player I misbehave, he will be in such a bad bargaining position that the outcome he receives will be inferior to the outcome if he behaves himself.

We can also derive sufficient conditions for renegotiation not to occur:

**THEOREM 4:** *Renegotiation will not occur if*

$$(9) \quad h - f \geq s - q,$$

$$(10) \quad \text{or } g - r \geq e - p,$$

$$(11) \quad \text{or } p + q > r + s.$$

Condition (9) makes it likely that an unrenegotiated liquidated damage contract will be successful. Condition (10) makes it likely that an unrenegotiated specific-performance contract will be successful. Finally, under condition (11), if player I misbehaves, there will be no renegotiation, and so no useful way to incorporate threats based on inferior bargaining positions.

The framework used in the law-and-economics literature typically restricts payoffs so that player II's payoffs in the game are unaffected by the unverifiable actions of player I. William Rogerson (1984), follows this pattern, so that in his model the payoffs are such that our equation (9) holds as an equality. Rogerson does not consider liquidated damage contracts in his framework. If they were included, they would dominate all the other forms of remedy, would never be renegotiated and would always achieve full-information efficient outcomes.

Thus although the possibility of renegotiation restricts the achievable outcomes in law-and-economics models, renegotiation for strategic purposes would never actually be observed in such models. Strategic renegotiation can only occur if each player's payoff is affected by the other's action.

#### E. Generalization to Multiple-Observable Actions

The procedure described above can be generalized in a straightforward way to handle the case when there are an arbitrary number of verifiable actions that player II can take. Let these actions be numbered  $x_0$  through  $x_n$ . If action  $i$  is taken and player I has played  $a$  then the payoffs to the two players are  $e_i$  and  $f_i$ , respectively. If  $b$  has been played, then the corresponding payoffs are  $p_i$  and  $q_i$ , respectively. Let  $(a, x_0)$  be the efficient pair of actions (so that the quartet  $(e_0, f_0, p_0, q_0)$  corresponds to the quartet  $(g, h, r, s)$  of the previous section). Then analysis exactly as before leads to the following theorem:

**THEOREM 5:** *If there is a verifiable act  $x_i$  such that  $e_i \geq p_i$  and  $q_i \geq f_i$ , then efficiency is achievable with a specific-performance contract.*

In other words, if given  $x_i$  player I prefers to behave and player II prefers player I not to behave, then efficiency can be achieved by committing player II to play  $x_i$ . If  $x_i \neq x_0$ , then the specific-performance contract will be subsequently renegotiated.

**THEOREM 6:** *If there exist two actions  $x_i$  and  $x_j$  such that (a)  $f_i \geq f_j$ ; (b)  $q_j \geq q_i$ ; (c)  $e_i \geq p_j$ ; and (d)  $q_j \geq f_i$ , then efficiency is achievable with a liquidated damages contract.*

In other words, if there exist two actions  $x_i$  and  $x_j$  such that (a) given I has behaved,  $x_i$  is better for II; (b) given I has misbehaved,  $x_j$  is better for II; (c) I prefers behaving plus  $x_i$  to misbehaving plus  $x_j$ ; and (d) II prefers the opposite, then the efficient outcome can be achieved. In this case we use a contract in which II can

choose either  $x_i$  or  $x_j$ . In effect,  $x_i$  is the action promised if  $a$  is played, and  $x_j$  is the action threatened if  $b$  is played. If  $x_i$  does not equal  $x_0$ , this contract will be renegotiated.

The following gives general necessary and sufficient conditions for achieving efficiency:

**THEOREM 7:** *Efficient outcomes can be achieved if and only if there exist two observable actions  $x_i$  and  $x_j$  and a number  $D$  such that the following inequalities are satisfied: (a)  $f_i - f_j \geq D$ ; (b)  $q_i - q_j \leq D$ ; and (c)  $D \leq c(e_0 + f_0 - \max_k (p_k + q_k)) - c(f_i - q_j) + (1 - c)(e_i - p_j)$ .*

If these inequalities are satisfied for some  $x_i$ ,  $x_j$ , and  $D$ , then a contract of the following form achieves the efficient outcome: If player II chooses action  $x_i$ , then he pays  $X$  to player I. If he chooses action  $x_j$ , he pays  $X + D$ . (If he chooses any other action he pays an infinite amount.)

This theorem can be understood as follows: Optimality requires the initial contract to include a threat coupled with a promise—the threat to be used only if player I misbehaves, and the promise only if player I behaves himself. If and only if both equations (a) and (b) above hold, then the marginal value of retaliation over nonretaliation is greater when player I misbehaves than when he behaves himself. In this case, side payments can be attached to  $x_i$  and  $x_j$  so that player II prefers retaliation *ex post* if and only if I has misbehaved.

Thus if equations (a) and (b) hold, there are side payments which make player II's threat credible *ex post*. For this threat to be effective, it must punish player I sufficiently to prevent him from misbehaving. This is the point of inequality (c) which guarantees player I prefers to make the socially desirable action, once he takes into account side payments and gains from bargaining in any renegotiation that will occur.

This theorem leads to general predictions as to when strategic renegotiation will be observed. First, if among the set of possible verifiable actions  $x_0$  is relatively desirable when player I has misbehaved, then renegotiation is likely to be observed in attain-

ing an optimum. For there will not be any threat against which  $x_0$  can serve as a credible promise; instead some other action must serve as the interim promise with subsequent renegotiation to the optimum.

Second, if the socially desirable pair  $(a, x_0)$  is unfavorable for player *I* and favorable for player *II*, (relative to other pairs  $(b, x_j)$  that can be played) renegotiation is likely to be necessary to achieve it. For  $x_0$  will not be an effective promise unless the outcome is relatively desirable to player *II*—so again some other interim promise must be used, to give player *II* a sufficient amount of surplus, with subsequent renegotiation to the efficient choice  $x_0$ .

A similar analysis can be applied when the number of unverifiable actions increases. Note that as the number of verifiable actions increases it becomes easier to achieve efficiency (because there are more potentially useful threats that can be included in a contract). As the number of unverifiable actions increases, it becomes more difficult to achieve efficiency since the number of potential defections increases.

### V. Concluding Remarks

In an intertemporal general equilibrium, no trade is desired after the first period if markets are complete. If, however, markets are incomplete, trade may be desired after the first period. Our account of contract renegotiation resembles these familiar results from general equilibrium theory. If all possible contracts could be written, no need for renegotiation would arise, even if the parties' relation is dynamic. Once we limit the universe of feasible contracts, renegotiation becomes desirable.

In order to explain contract renegotiation it seems necessary to posit a cost to writing complicated contracts. For with no limit to the complexity that can be entailed, anything which can be achieved through renegotiation can be achieved by a complicated contract with a sequence of announcements: the individuals mirror the renegotiation process through announcements, where the interpretations to be attached to

various announcements are prespecified by the contract.

That any renegotiation process can be emulated by writing elaborate contracts which are unrenegotiated does not vitiate the paper's main point. The elaborate contracts are not superior to contract renegotiation; they merely replicate it. In the future we hope to model explicitly the costs of complex contracts and of contract negotiations in order to identify circumstances in which renegotiation is preferred to writing a complex contract.

This paper demonstrates that once unverifiable actions are present in a circumstance which calls for a contract, then contract renegotiation becomes a useful tool even in the absence of objective uncertainty. As the values of unverifiable parameters are realized, the contracting parties' preferences over the feasible set of contracts changes. The result is a natural role for limited contracts, social preference not to contract, and renegotiation.

In particular, we will then observe contracts which, rather than being renegotiated only when unexpected contingencies arise, will be renegotiated with certainty and with both parties fully anticipating this possibility. We have characterized the cases in which this result occurs, noting that it is necessary that each player have a tendency to respond noncooperatively when the other cooperates.

### REFERENCES

- Baldwin, Carliss, "Productivity and Labor Unions: An Application of the Theory of Self-Enforcing Contracts," *Journal of Business*, April 1983, 56, 155–86.
- Binmore, Kenneth G., "Perfect Equilibrium in Bargaining Models," ICERD Discussion Paper No. 58, London School of Economics, 1982.
- \_\_\_\_\_, Rubinstein, Ariel, and Wolinsky Asher, "The Nash Bargaining Solution in Economic Modelling," *Rand Journal of Economics*, Summer 1986, 17, 176–88.
- Crawford, Vincent P., "Long-Term Relationships Governed by Short-Term Contracts," Harvard Institute of Economic Research, Discussion Paper No. 926, 1982.

- Dye, Ronald A., "Optimal Length of Labor Contracts," *International Economic Review*, February 1985, 26, 251-70.
- Dybvig, Philip and Spatt, Chester, "Does it Pay to Maintain a Reputation? Quality Incentives in Financial Markets," mimeo., 1985.
- Grossman, Sanford J. and Hart, Oliver D., "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," *Journal of Political Economy*, August 1986, 94, 691-719.
- Hart, Oliver D. and Moore, John, "Incomplete Contracts and Renegotiations," mimeo., 1985.
- Harris, Milton and Holmstrom, Bengt, "On the Duration of Agreements," *International Economic Review*, June 1987, 28, 389-405.
- Huberman, Gur and Kahn, Charles M., "On the Scope of Contracts and Contract Renegotiations," unpublished Working Paper, University of Chicago, 1985.
- \_\_\_\_\_, and \_\_\_\_\_, "Secured Loans, Default, and Strategic Renegotiation," paper presented at the *Conference on the Economics of Contract Law*, Duke University, Durham, NC, April 8-9 1988.
- Klein, Benjamin and Leffler, Keith B., "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, August 1981, 89, 615-41.
- Leland, Hayne E. and Pyle, David H., "Informational Asymmetries Financial Structure, and Financial Intermediation," *Journal of Finance*, May 1977, 32, 371-87.
- Mnookin, Robert H. and Kornhauser, Lewis, "Bargaining Under the Shadow of the Law: The Case of Divorce," *Yale Law Journal*, April 1979, 88, 950-97.
- Rogerson, William P., "Efficient Reliance and Damage Measures for Breach of Contract," *Rand Journal of Economics*, Spring 1984, 15, 39-53.
- Rubinstein, Ariel, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, January 1982, 50, 97-109.
- Shapiro, Carl, "Consumer Information, Product Quality, and Seller Reputation," *Bell Journal of Economics*, Spring 1982, 13, 20-35.
- Shavell, Steven, "Damage Measures for Breach of Contract," *Bell Journal of Economics*, Autumn 1980, 11, 466-90.
- \_\_\_\_\_, "The Design of Contracts and Remedies for Breach," *Quarterly Journal of Economics*, February 1984, 99, 121-48.
- Tirole, Jean, "Procurement and Renegotiation," *Journal of Political Economy*, April 1986, 94, 235-59.
- Williamson, Oliver E., "Credible Commitments: Using Hostages to Support Exchange," *American Economic Review*, September 1983, 73, 519-40.

# Long-Term Relationships Governed by Short-Term Contracts

By VINCENT P. CRAWFORD\*

*This paper studies the effect of contract duration on the incentive to invest in a relationship when its parties are rational and have perfect information, and contracts are complete, except that short-term contracts specify only current-period actions. Then, short-term contracting distorts investment decisions only when the efficient plan involves mainly sunk-cost investment and the relationship plays a consumption-smoothing role. There is a general, but not universal, tendency to underinvest.*

Contractual relationships usually outlive the contracts that govern them. Long-term relationships are therefore normally governed by sequences of short-term contracts. When these allow complete control over a relationship's organization within each contract period, any plan that would be feasible with a complete long-term contract, covering the relationship's entire life, is also feasible with a sequence of short-term contracts.

This fact may make long-term commitments appear unimportant, particularly when the parties to the relationship have perfect foresight. The duration of contracts can nevertheless have significant effects, because short-term contracts must be voluntarily negotiated in the bargaining environments created by earlier contracts. This places constraints on the agreements rational parties can reach that a long-term contract would allow them to override.

Because short-term contracting is so prevalent, it is worthwhile to ask when these

constraints cause inefficiency and what form the distortion takes.<sup>1</sup> This paper studies these questions for relationships in which efficiency requires investment in *relationship-specific capital*—that is, capital whose productivity depends on the continuation of the relationship. In order to focus on the interaction between short-term contracting and relationship-specific investment, I assume that parties have perfect information and perfect foresight; that they are rational, and use their contracting possibilities efficiently; and that contracts are costless to enforce and complete, except that short-term contracts do not allow commitments to actions taken beyond the contract period. To maintain control over the duration of agreements, I also rule out “implicit” contracts.

It is useful to distinguish between two polar types of relationship-specific investment: *reversible* investment, which creates capital that can later be withdrawn from the relationship and either consumed or reinvested elsewhere, and *sunk-cost* investment, which creates capital that cannot be withdrawn. It is shown below, under the assumptions described above, that whether short-term contracting distorts relationship-specific investment is determined by two factors:

\*Department of Economics D-008, University of California, San Diego, La Jolla, CA, 92093. I am grateful to the many people who made helpful comments on earlier versions of this paper, especially Peter Berck, Zoë Crawford, Milton Harris, Bengt Holmstrom, George Jakubson, Mark Machina, James Mirrlees, Dennis Smallwood, Joel Sobel, Jean Tirole, and a referee. Partial financial support was provided by the Industrial Relations Section, Princeton University, and by the National Science Foundation under grants SES-8204038, SES-8408059, and SES-8703337. The continuous-time analysis in Section III of my 1986 paper bearing this title will be published separately, under the title “Relationship-Specific Investment.”

<sup>1</sup>I do not consider the intriguing related question of why the contracts we observe have a special tendency to be incomplete intertemporally. This could be explained by a model in which bargaining requires a setup cost, adding contingencies to a contract is costly, and uncertainty about which future contingencies remain relevant is resolved gradually over time.

parties' need to smooth consumption over time within their relationship, and the predominance of sunk cost over reversible investment in the efficient plan.

Gary Becker's (1962) analysis of a firm's incentive to train its workers in general skills illustrates the importance of parties' need to use their relationship for consumption smoothing. He noted that workers cannot make effective agreements limiting their future wages, because labor law does not allow them to bind themselves irrevocably to their employers. It follows that once a worker has acquired skills of general applicability, competition will bid up his wages until they fully reflect his increased productivity. Thus, a firm cannot rationally expect to capture any of the future returns from training its workers in general skills.

Why, then, do firms undertake such investments? Becker gave an elegant answer: If a worker has access to a perfect capital market, he can costlessly compensate his firm for its investment in training him by agreeing to accept lower wages under the current contract. Even when long-term commitments cannot be enforced, this allows parties to realize the benefits of efficient investment that are usually presumed to follow from long-term contracting. This result is generalized in Section I. There it is shown, under the assumptions outlined above, that when parties do not need their relationship for efficient consumption smoothing—and, in particular, when they have access to perfect capital markets—short-term contracting yields the same efficient outcome as long-term contracting.<sup>2</sup>

Section I's result provides a point of departure for my analysis, which is motivated

by two observations. First, Becker's analysis of investment incentives greatly overstates the ease with which the problems he studied are solved in practice. Paying for his own training would often require a worker to finance his consumption for several years by borrowing in a highly imperfect capital market. It may still be *feasible* for him to provide sufficient incentives for efficient investment by compensating his firm under the current contract, but this can induce costly consumption distortions. Short-term contracting then requires a second-best compromise between efficient investment and an efficient time pattern of consumption; such compromises are important determinants of resource allocation in many relationships.

Second, many economists feel intuitively that sunk costs have a special tendency to distort investment decisions. Constructing a theory with which to evaluate this intuition is one of my main goals. When (as in the models studied below) parties never wish to withdraw reversible capital, sunk-cost and reversible investment differ solely in how they affect parties' alternatives to remaining in their relationship. Because these effects come into play only after they have agreed on how much to invest, long-term contracting makes sunk-cost and reversible investment equivalent. It follows from the analysis of Section I that they are also equivalent, under short-term contracting, when parties do not need their relationship for efficient consumption smoothing. Thus, in my framework, sunk costs can cause investment distortions only when short-term contracting interferes with consumption smoothing within the relationship.

With these observations in mind, Sections II and III study parties' incentives to invest in their relationship when their preferences over consumption streams are representable by concave, additively separable utility functions and they cannot borrow or lend outside the relationship. Section II follows most of the literature in assuming that one party (the same one in each contract period) makes all-or-nothing contract proposals, which the other party can only accept or reject. In perfect equilibrium in this highly simplified model of bargaining, the party who makes

<sup>2</sup>The risk-neutrality and "wealth-maximization" assumptions commonly maintained in the literature on law and economics (see, for example, Benjamin Klein, Robert Crawford, and Armen Alchian, 1978, and Oliver Williamson, 1979, 1983) also imply that, when assets can costlessly be transferred, parties do not need their relationship for efficient consumption smoothing. It is likely that the problems considered here have not been modeled explicitly in that literature because they cannot arise under these assumptions with symmetric information.



proposals extracts all of the surplus from the relationship, limited only by the contracting possibilities and the other party's right to reject proposals. Rationality therefore requires him to structure his proposals to maximize the present and anticipated future surplus subject to these constraints, ensuring that, as in Section I, parties use their contracting possibilities efficiently. Section III tests the robustness of Section II's results by studying symmetric surplus sharing, with contracts determined by John Nash's (1950, 1953) bargaining solution.

In each case, a party who makes a sunk-cost, relationship-specific investment without the protection of a long-term contract must be compensated under the current contract, or not at all. For, who made such an investment has no effect on the bargaining environments in which future contracts will be negotiated; in the absence of implicit contracts, it therefore does not affect the bargains struck in those environments. (The investor may of course share in the enhanced productivity of the relationship, but no more than if his partner had made the investment.) To the extent that the need for consumption smoothing makes current compensation costly, this constraint on compensation for sunk-cost investment makes the investor like a private provider of a public good, who pays full cost but receives only part of the benefit. This analogy suggests that short-term contracting normally leads to inefficiently low levels of sunk-cost investment; the analysis confirms this, but also shows that parties sometimes overinvest in order to commit themselves to a more efficient time pattern of consumption.

By contrast, the level of reversible investment is normally efficient even under short-term contracting. The crucial difference between reversible and sunk-cost investment is that only reversible investment creates capital whose value outside the relationship is commensurate with that of the consumption it generates within the relationship. If the efficient plan creates enough reversible capital and the efficient time pattern of consumption is "normal" (in a sense made precise below), each short-term contract can specify ownership of the reversible capital so that

parties will voluntarily negotiate the efficient pattern of consumption in subsequent contracts, restoring the incentives for efficient investment. Although it is also possible to assign ownership of sunk-cost capital, it has no value outside the relationship and therefore cannot serve as a substitute for commitments about future compensation.

Section IV concludes the paper with a discussion of related work.

### I. Short-Term Contracting Without Consumption Smoothing

This section formalizes and extends Becker's (1962) observation that parties who have access to perfect capital markets may not need long-term contracts to ensure efficient investment in their relationship. Under the assumptions outlined above, it is shown that when parties do not need their relationship for efficient consumption smoothing, short-term contracting is equivalent to long-term contracting, so that the duration of contracts affects neither the efficiency of their relationship's organization nor how they share its surplus. This motivates the contrasting assumptions about the need for consumption smoothing used below to evaluate the effect of contract duration on relationship-specific investment.

There are two parties, called  $U$  and  $V$ , who have perfect information and perfect foresight. They can cooperate in production in each of a finite number of periods. In each period in which they cooperate, their efforts produce an output of a divisible consumption good; they set aside some of this output as an investment in their relationship and share the rest. In periods in which they do not cooperate, their individual efforts produce *autarkic* outputs that are independent of how the relationship is organized.

The central assumption of this section's analysis is that  $U$  and  $V$  do not need their relationship for efficient consumption smoothing, in the following sense: Given any investment plan that is compatible with efficiency and any distribution of the resulting stream of output, they can achieve an efficient time pattern of consumption without trading with each other. To make this idea

precise, call a party's share of a given period's output his *compensation*, to distinguish it from his consumption when they differ. Assume that  $U$ 's and  $V$ 's preferences over time paths of compensation when there are  $t$  periods to go in their relationship are represented by weakly concave utility functions  $U_t(\cdot)$  and  $V_t(\cdot)$ , which reflect whatever opportunities they have for borrowing and lending outside their relationship as well as their underlying preferences for consumption smoothing. Then  $U$  and  $V$  do not need their relationship for efficient consumption smoothing if and only if their preferences over compensation streams have indifference surfaces that are coincident hyperplanes. It is well known (see, for example, Andreu Mas-Colell, 1985, pp. 80–81) that if all of the level surfaces of a concave function are hyperplanes, they must be parallel. Thus, when  $U$  and  $V$  do not need their relationship for efficient consumption smoothing, their preferences can be represented by identical, linear utility functions.<sup>3</sup>

$U$  and  $V$  negotiate a new contract at the start of each period not covered by an existing contract. (Under my assumptions, it is never mutually beneficial to renegotiate a contract.) Contracts are costless to enforce and complete, except that short-term contracts do not allow commitments to actions taken beyond the contract period. A complete long-term contract specifies investment, compensation, and whether parties cooperate in production for each period that remains in their relationship. A complete short-term contract specifies the same variables, but only for the current period. Finally,  $U$ 's and  $V$ 's contract agreements are determined by an efficient bargaining solution that depends only on the utility combinations that are feasible for their relationship, given the remaining time horizon, and the utilities they obtain if they do not agree in the current period. (This is consistent with the assumptions about bargaining outcomes

maintained in Sections II and III.) The result can now be stated:

**PROPOSITION 1:** *When parties do not need their relationship for efficient consumption-smoothing, short-term contracting yields the same efficient outcome as long-term contracting. In each case, bargaining shares the surplus from the relationship as if disagreement precluded cooperation for the rest of its life.*

#### PROOF:

It is immediate, under my assumptions, that long-term contracting ensures efficient organization of the relationship. To see that short-term contracting also does this, note that when  $U$  and  $V$  have identical, linear preferences over compensation streams, they can adjust their compensations in any period to transfer utility between them at a constant, one-for-one rate. This implies that a (long- or short-term) contract with  $t$  periods to go is efficient, if and only if it maximizes  $U_t(\cdot) + V_t(\cdot)$ , given  $U$ 's and  $V$ 's common rational expectations of how future contracts are influenced by the current contract. Let  $u_t$  and  $v_t$  stand for  $U$ 's and  $V$ 's compensations in period  $t$  and let  $a$  stand for their common discount factor (assumed constant over time for simplicity only). Summing the resulting expressions for  $U_t(\cdot)$  and  $V_t(\cdot)$  yields

$$(1) \quad U_t(\cdot) + V_t(\cdot) \equiv \sum_{i=1}^t a^{t-i}(u_i + v_i);$$

an efficient long-term contract specifies parties' decisions in each period to maximize this expression over all feasible plans. Because this problem has time-consistent solutions, a sequence of efficient short-term contracts, each of which maximizes the sum of parties' utilities for the rest of their relationship given their expectations that future contracts will also do this, must also yield decisions that maximize the expression in (1) over all feasible plans. Short-term contracting therefore ensures efficient organization of the relationship as well.

It remains to be shown that parties share the surplus from their relationship in the same way under long- and short-term con-

<sup>3</sup>With perfect capital markets, for instance, parties' utilities are determined by the present values of their compensation streams.

tracting. Because parties' contracting possibilities within the relationship do not influence their autarkic outputs, this is immediate when disagreement in any period precludes cooperation for the rest of their relationship. The proof given here assumes, instead, that disagreement precludes cooperation only in the period in which it occurs; the proof for longer interruptions differs only in notation.

An efficient plan for the relationship determines, for any period, the total surplus  $U$  and  $V$  can share over the rest of its life; let  $W_t$  denote this total when there are  $t$  periods to go. Also for period  $t$ , let  $u_t$  and  $v_t$  denote  $U$ 's and  $V$ 's autarkic outputs;  $\underline{U}_t \equiv \sum_{i=1}^t a^{t-i} u_i$  and  $\underline{V}_t \equiv \sum_{i=1}^t a^{t-i} v_i$  their autarkic utilities for the rest of the horizon; and  $U_t^d$  and  $V_t^d$  their utilities if they do not make a contract. Suppose that the bargaining solution shares the surplus from each contract agreement between  $U$  and  $V$  in the proportions  $\alpha: 1-\alpha$ , with  $0 \leq \alpha \leq 1$ .

I now argue that long- and short-term contracting both yield the same division of surplus as would bargaining when disagreement precludes cooperation for the rest of the relationship's life. The proof is the same (with appropriate definitions of the variables) in each case; it proceeds by induction. The desired conclusion is immediate for the last period. Assume that it is true for the last  $t-1$  periods. Bargaining in period  $t$  then yields  $U$  and  $V$  utilities

$$(2) \quad U_t = U_t^d + \alpha(W_t - U_t^d - V_t^d) \\ = \alpha W_t + (1-\alpha)U_t^d - \alpha V_t^d$$

and

$$(3) \quad V_t = V_t^d + (1-\alpha)(W_t - U_t^d - V_t^d) \\ = (1-\alpha)W_t - (1-\alpha)U_t^d + \alpha V_t^d,$$

where  $U_t^d$  and  $V_t^d$  are given by

$$(4) \quad U_t^d = \underline{u}_t + a\underline{U}_{t-1} = \underline{u}_t \\ + a[\alpha W_{t-1} + (1-\alpha)\underline{U}_{t-1} - \alpha \underline{V}_{t-1}]$$

and

$$(5) \quad V_t^d = \underline{v}_t + aV_{t-1} = \underline{v}_t \\ + a[(1-\alpha)W_{t-1} - (1-\alpha)\underline{U}_{t-1} + \alpha \underline{V}_{t-1}].$$

The right-hand equalities in (4) and (5) follow from the induction hypothesis and equations (2) and (3) for period  $t-1$ . Substituting (4) and (5) into (2) yields

$$(6) \quad U_t = \alpha W_t + (1-\alpha)\{\underline{u}_t + a \\ \times [\alpha W_{t-1} + (1-\alpha)\underline{U}_{t-1} - \alpha \underline{V}_{t-1}]\} \\ - \alpha\{\underline{v}_t + a[(1-\alpha)W_{t-1} \\ - (1-\alpha)\underline{U}_{t-1} + \alpha \underline{V}_{t-1}]\} \\ = \alpha W_t + (1-\alpha)(\underline{u}_t + a\underline{U}_{t-1}) \\ - \alpha(\underline{v}_t + a\underline{V}_{t-1}) \\ = \alpha W_t + (1-\alpha)\underline{U}_t - \alpha \underline{V}_t;$$

the analogous expression for  $V$  follows immediately from efficiency, establishing the desired conclusion for period  $t$ .

Proposition 1's efficiency conclusion stems from the fact that, when current compensation for the future effects of current decisions is costless, efficient short-term contracting internalizes these effects, ensuring that parties maximize the surplus from their relationship over its entire life. It follows that, under the maintained assumptions, any effects of contract duration on the level of investment in the relationship must stem from interactions between investment and parties' need to smooth consumption within their relationship.<sup>4</sup>

<sup>4</sup>It bears emphasis that this conclusion may not hold without my assumptions that information is symmetric and bargaining is efficient, and that contracts are complete in the short run. However, Drew Fudenberg, Bengt Holmstrom, and Paul Milgrom (1987) have proved a result analogous to Proposition 1 for a class of long-term agency relationships with asymmetric information and moral hazard.

The assumptions maintained in Sections II and III are designed to introduce such interactions into the analysis in a reasonably general, but controlled way.

## II. Investment Incentives with All-or-Nothing Contract Proposals

This section begins to study how the duration of contracts affects parties' incentives to invest in their relationship. Its model is a special case of Section I's, with more physical detail and a simple, tractable characterization of bargaining outcomes. As in Section I, there are two parties, called  $U$  and  $V$ , who are perfectly informed and rational, in the standard sense that they play perfect-equilibrium (henceforth shortened to "equilibrium") strategies in the bargaining games specified below. Contracts remain costless to enforce and complete, except that short-term contracts do not allow commitments to actions taken beyond the contract period.

There are two periods, numbered 1 and 2.  $U$  and  $V$  can cooperate in production in either or both periods. Their efforts produce an output of a divisible consumption good, which is costless to store, or to transfer between them.  $U$  and  $V$  cannot borrow or lend outside their relationship, but they can set aside some of the first period's output as an investment in their relationship. (Investment is unproductive until the period after it is made, so in this model it is pointless to invest in the second period.) Total output of the consumption good is higher, other things equal, when  $U$  and  $V$  cooperate in production; and total cooperative output is higher, the more they have invested.

Let  $z(s)$  denote the output of the relationship in any period in which  $U$  and  $V$  cooperate in production, where  $s$  represents the amount of relationship-specific investment made by the start of the period. Let  $x$  and  $y$  denote  $U$ 's and  $V$ 's strictly positive autarkic outputs in any period in which they do not cooperate;  $x$  and  $y$  are independent of  $s$  by definition.<sup>5</sup> The assumptions just described

imply that  $z(s)$  is increasing in  $s$ , and that  $z(0) > x + y$ . It is also assumed that  $z(\cdot)$  is strictly concave and continuously differentiable, with  $z'(0) = \infty$  to ensure that parties always undertake some investment in equilibrium.

$U$  and  $V$  have additively separable preferences over consumption streams with stationary felicity functions  $u(\cdot)$  and  $v(\cdot)$ , which are increasing, strictly concave, and continuously differentiable, with  $u'(0) = v'(0) = \infty$ . Thus,  $U$ 's and  $V$ 's utility functions are  $u(c_1) + au(c_2)$  and  $v(d_1) + bv(d_2)$ , respectively, where  $c_i$  and  $d_i$  denote  $U$ 's and  $V$ 's consumptions in period  $i$  and the discount factors  $a$  and  $b$  are strictly positive and less than or equal to one.

In this section, the rules of bargaining are as follows. At the start of each period that is not covered by an already existing contract,  $U$  makes a complete, all-or-nothing contract proposal.  $V$  then accepts or rejects  $U$ 's proposal. Acceptance turns the proposal into a binding contract; rejection leads to autarky for an exogenous number of periods, either one or two. If time remains in the horizon after a contract expires or an autarkic interval ends,  $U$  begins the process again with a new proposal, and bargaining continues as before.

A complete long-term contract proposal specifies all relevant variables for the two-period horizon, or what remains of it. In the present model, these are: whether parties cooperate in production in each period, their consumption in each period, investment in the first period, and (sometimes) the assignment of ownership of the resulting capital. A complete short-term contract proposal specifies the same variables, but only for the current period.

The analysis rests on two observations about equilibrium strategies, which hold, in

<sup>5</sup>Gregory Dow (1985) and my paper (1983, Sec. IV) endogenize parties' autarkic outputs and study their

incentives, under short-term contracting, to take actions that shift them in ways that enhance their future bargaining power. An interesting implication of the present analysis is that having too much future bargaining power may pose problems for a party, in that under short-term contracting, it may require him to compensate his partner inefficiently early.

this environment, both for sunk-cost and reversible investment and for long- and short-term contracting.<sup>6</sup> First,  $V$  cannot reject any of  $U$ 's equilibrium proposals, because  $U$  could then do better by adjusting them to ensure that  $V$  accepts. Further, unless  $V$  is indifferent between accepting and rejecting a proposal,  $U$  can adjust it to extract more surplus. Thus,  $V$  is indifferent between accepting or rejecting  $U$ 's proposals in equilibrium, but always accepts them.

Second, if  $V$  rejected  $U$ 's first-period proposal, his utility would be  $(1+b)v(y)$ , his autarkic utility for the two-period horizon. Because  $V$  can never benefit from storage or investment on his own, this is immediate when rejection enforces autarky for two periods. Even when rejection enforces autarky for only one period,  $V$ 's second-period autarkic consumption following a first-period rejection is  $y$ ;  $U$  will therefore offer him  $y$  in his equilibrium second-period proposal. Thus,  $V$ 's utility is  $(1+b)v(y)$  in either case.

In equilibrium under long-term contracting,  $U$ 's first-period contract proposal specifies levels of investment and consumption that solve

$$(L) \quad \max_{s, d_1, d_2 \geq 0} [u(z(0) - s - d_1) + a \times u(z(s) - d_2)]$$

$$\text{s.t. } v(d_1) + bv(d_2) \geq (1+b)v(y),$$

and  $V$  accepts this proposal. To see this, note that solving (L) yields an efficient plan for the relationship that gives  $U$  the highest utility consistent with  $V$ 's right to reject proposals. Thus, if  $U$ 's first-period proposal specifies investment and consumption as determined by (L) and  $V$  accepts, parties will never agree to renegotiate the resulting contract. It then follows from the observations made above that  $U$ 's first-period proposal is consistent with equilibrium if and only if it

solves (L), and that  $V$  must accept this proposal in equilibrium; the resulting long-term contract fully determines parties' welfares. This argument does not depend on how long rejection enforces autarky, whether  $s$  represents sunk-cost or reversible investment, or how the ownership of the relationship's capital is assigned: in this case these factors are irrelevant.

My assumptions on preferences and technology ensure that (L) has a unique, interior solution, characterized by its constraint holding with equality and the first-order conditions

$$(7) \quad z'(s) = \frac{u'(z(0) - s - d_1)}{au'(z(s) - d_2)} = \frac{v'(d_1)}{bv'(d_2)}.$$

The efficiency conditions in (7) have a standard marginal rate of transformation equals marginal rates of substitution interpretation. Denote the values of  $s$ ,  $d_1$ , and  $d_2$  that solve problem (L) and satisfy (7)  $s^*$ ,  $d_1^*$ , and  $d_2^*$ ; they will serve as an efficient benchmark to compare with the short-term contracting outcome.

Under short-term contracting, the difference between sunk-cost and reversible investment matters; efficiency typically requires some of each. Such combinations are best understood by studying each of the two polar cases in turn.

When efficient organization of the relationship requires only sunk-cost investment,  $U$ 's equilibrium first-period contract proposal is found by solving problem (L) with the added constraints that  $d_1 = d_2 = y$ . To see this, note that  $V$ 's second-period consumption must be at least  $y$  for him to accept  $U$ 's proposal.  $U$  cannot commit himself to pay  $V$  more than  $y$  by assigning him ownership of the relationship's capital, because it has no value outside the relationship in this case; and it is never beneficial to use storage for this purpose. In equilibrium, therefore,  $V$ 's second-period consumption must be exactly  $y$ .  $U$ 's first-period contract proposal must maximize his utility, given his rational expectation that  $V$  will accept his equilibrium second-period proposal. The de-

<sup>6</sup>More precisely, the observations hold for  $U$ 's and  $V$ 's perfect-equilibrium strategies, both on and off the equilibrium path.

sired conclusion then follows immediately from the above observations.

Substituting the constraints  $d_1 = d_2 = y$  into (L), and noting that they make its constraint on  $V$ 's utility redundant, yields

$$(S) \max_{s \geq 0} u(z(0) - s - y) + au(z(s) - y).$$

Problem (S) determines the level of sunk-cost investment as a second-best response to the "perfectness" constraint on consumption smoothing associated with short-term contracting. My assumptions on preferences and technology ensure that (S) has a unique, interior solution, characterized by the first-order condition

$$(8) \quad z'(s) = \frac{u'(z(0) - s - y)}{au'(z(s) - y)}.$$

The value of  $s$  that solves (S) and satisfies (8) is denoted  $\hat{s}$ .

It is clear from (7) and (8) that  $\hat{s}$  generally differs from  $s^*$ , because  $s^*$  depends on the form of  $v(\cdot)$  but  $\hat{s}$  does not. I now consider when  $\hat{s} < s^*$ , as suggested by the private provision of public goods analogy discussed in the introduction. Although it relates endogenous variables, the following necessary and sufficient condition for  $\hat{s}$  to be less than  $s^*$  is useful in discerning what the difference between them depends on:

**LEMMA 1:**  $\hat{s} < s^*$  if and only if  $d_1^* < y < d_2^*$ .

**PROOF:**

It is clear that  $d_1^* < y$  if and only if  $d_2^* > y$  from the constraint of (L), which must hold with equality at its solution. To establish the "if" part of the conclusion, suppose that  $d_1^* < y < d_2^*$  but  $\hat{s} \geq s^*$ . Combining (7) and (8) and using the concavity of  $z(\cdot)$  then yields

$$(9) \quad z'(\hat{s}) = \frac{u'(z(0) - \hat{s} - y)}{au'(z(\hat{s}) - y)} \leq \frac{u'(z(0) - s^* - d_1^*)}{au'(z(s^*) - d_2^*)} = z'(s^*) < \frac{u'(z(0) - s^* - d_1^*)}{au'(z(s^*) - d_2^*)}.$$

But, given the binding constraint of (L) and the concavity of  $u(\cdot)$ , the inequality in (9) implies that  $z(0) - \hat{s} - y \geq z(0) - s^* - d_1^*$ . Together with the hypothesis that  $\hat{s} \geq s^*$ , this implies that  $y \leq d_1^*$ , a contradiction. To prove the "only if" part, suppose that  $d_1^* \geq y \geq d_2^*$  but  $\hat{s} < s^*$ . Then  $z(0) - s^* - d_1^* < z(0) - \hat{s} - y$  and, given the binding constraint of (L),  $z(s^*) - d_2^* > z(\hat{s}) - y$ . It then follows from (7) and (8), given the strict concavity of  $u(\cdot)$ , that  $z'(s^*) > z'(\hat{s})$ , contradicting the concavity of  $z(\cdot)$  and the hypothesis that  $\hat{s} < s^*$ .

The intuition behind Lemma 1 is simple. Under short-term contracting, the level of sunk-cost investment affects the time pattern of consumption. If  $d_1^* < y < d_2^*$ , reducing investment below  $s^*$  brings the relationship between parties' consumption streams closer to the efficient pattern. Parties then invest less than productivity considerations alone would dictate, in exchange for an improved consumption pattern. If, instead,  $d_1^* > y > d_2^*$ , it is overinvestment that makes the pattern of consumption more efficient.

To see that Lemma 1's condition for short-term contracting to lead to too little sunk-cost investment represents the "normal" case, think of  $U$  and  $V$  as traders in a pure-exchange economy, with endowments  $(z(0) - s^*, z(s^*))$  and  $(y, y)$ . Because  $U$  and  $V$  are averse to time variation in consumption, efficient trade normally brings the slopes of their time paths of consumption closer together. As  $z(0) - s^* < z(s^*)$ , this implies that  $d_1^* < y < d_2^*$ . The next result establishes a more precise sense in which too little sunk-cost investment is the norm:

**LEMMA 2:** If  $a \leq b$ , then  $d_1^* < y < d_2^*$ , and therefore  $\hat{s} < s^*$ .

**PROOF:**

Suppose, by way of contradiction, that  $a \leq b$  but  $d_1^* \geq y \geq d_2^*$ . Then

$$(10) \quad \frac{v'(d_1^*)}{bv'(d_2^*)} \leq \frac{1}{b} \leq \frac{1}{a} < \frac{u'(z(0) - s^* - d_1^*)}{au'(z(s^*) - d_2^*)}.$$

where the inequalities follow from the strict concavity of  $u(\cdot)$  and  $v(\cdot)$  and the fact that  $s^* > 0$ . The strict inequality in (10) contradicts the right-hand first-order condition in (7), and therefore establishes that  $d_1^* < y < d_2^*$ . That  $\hat{s} < s^*$  then follows immediately from Lemma 1.

Lemma 2's sufficient condition for  $\hat{s} < s^*$  is far from necessary, but it is evident from problem (L) and Lemma 1 that it is always possible for  $b$  to fall far enough below  $a$  for  $\hat{s}$  to exceed  $s^*$ . In such cases, parties agree to overinvest, despite the private provision of public goods intuition, in order to commit to a more efficient pattern of consumption. The effect of short-term contracting on the level of sunk-cost investment can be summarized as follows:

**PROPOSITION 2:** *When efficient organization of the relationship requires only sunk-cost investment, short-term contracting with all-or-nothing offers normally yields an inefficiently low level of investment in the relationship.*

A similar conclusion holds in Becker's (1962) model of a firm's decision to invest in training its workers in generally applicable skills; in my paper (1983), Section IV provides a formal analysis. It is clear that in the present model, a long-term commitment by  $U$  (whose role is traditionally the firm's in labor-contracting models) would normally suffice for efficient investment. This is not true in Becker's model (where the worker would have to make a long-term commitment) or in the symmetric-bargaining analysis of Section III.

Now consider the opposite polar case, in which efficient organization of the relationship requires only reversible investment. Recall that reversible investment is defined by the absence of sunk costs, so that the owner of the capital it creates can costlessly withdraw and either consume or reinvest it outside the relationship. I now show that, within limits set by the value of the reversible capital outside the relationship, its ownership can be assigned in short-term contracts to ensure that parties voluntarily

negotiate the pattern of compensation that supports efficient relationship-specific investment.

Assume, using the same notation as before, that the solution of (L) satisfies  $d_1^* \leq y \leq d_2^* \leq y + s^*$ . Let  $U$ 's first-period contract proposal specify  $s = s^*$  and  $d_1 = d_1^*$ , and assign  $V$  ownership of  $d_2^* - y \leq s^*$  units of the capital created by the investment. Ownership allows  $V$  costlessly to withdraw his capital from the relationship if he does not accept  $U$ 's second-period proposal.<sup>7</sup> This in effect raises his second-period autarkic consumption from  $y$  to  $y + (d_2^* - y) = d_2^*$ , ensuring that his second-period consumption will be  $d_2^*$ , and that he will accept  $U$ 's first-period proposal. Thus, assigning ownership of reversible capital allows  $U$ , with  $V$  responding optimally, to achieve the long-term contracting outcome with short-term contracts. Because the long-term contracting outcome is efficient and gives  $V$  his equilibrium utility under short-term contracting,  $U$  cannot improve upon it. These strategies are therefore in equilibrium.<sup>8</sup>

The conditions that  $d_1^* \leq y \leq d_2^*$  and  $d_2^* \leq y + s^*$  are both restrictive and necessary for this conclusion. When efficient consumption smoothing does not fit the normal pattern, efficiency would require  $V$  to commit himself to consume  $d_2^* < y$  in the second period, and there is no way to do this

<sup>7</sup>The enforceability of property rights over horizons longer than the current contract assumed here is not incompatible, logically or realistically, with my assumption that parties cannot make long-term contracts. Some enforceability of this kind is implicit in the assumption, universally maintained even when long-term contracts are not allowed, that a party cannot be forced to cooperate.

<sup>8</sup>Nothing in this argument depends on using property rights generated within the relationship by reversible investment: any assets that can be costlessly transferred and have value outside the relationship would serve as well. To the extent that parties begin their relationship with such assets, the distortions associated with short-term contracting are smaller than my analysis suggests. However, parties rarely have enough costlessly transferable assets to eliminate the investment distortions I study. Assuming that parties have assets that are costly to transfer implies that they have preferences over compensation streams different from those assumed in the text, but with similar implications.

with short-term contracts. Interestingly, when there is too little reversible capital to restore efficiency, short-term contracting causes investment distortions that resemble, but are smaller than, those that arise in the pure sunk-cost case, contrary to popular intuition. This conclusion extends to the general case where efficiency requires both sunk-cost and reversible investment, possibly technologically linked. The above results can be summarized as follows:

**PROPOSITION 3:** *When efficient organization of the relationship requires enough reversible investment, short-term contracting with all-or-nothing offers normally ensures efficient investment in the relationship. Otherwise, short-term contracting causes investment distortions qualitatively similar to those that arise when efficiency requires only sunk-cost investment, but smaller.*

Because sunk-cost and reversible investment have such different implications, it is worth emphasizing that the essential difference between them lies not in the possibility of assigning ownership of the capital they create, but in its value (net of the cost, if any, of enforcing property rights) outside the relationship. Only reversible investment creates capital whose value outside the relationship is commensurate with its consumption benefits within the relationship, and therefore comparable to the value required to enable parties to ensure efficient consumption smoothing by assigning its ownership. Sunk-cost capital has no value outside the relationship, so owning it cannot substitute for contractual commitments about future compensation.

Finally, although I have considered parties' relationship in isolation, summarizing the effects of the outside world by their autarkic outputs, the analysis also applies when the relationship has competition. Parties' autarkic outputs can be thought of as what they could obtain either working for themselves or in their best alternative relationships, which may face similar contracting problems. In this interpretation, my results show that when efficiency requires mainly reversible investment, sophisticated

use of short-term contracts can allow parties to make competition do the work of a long-term contract; this confirms—with some qualifications—a popular intuition. When efficiency instead requires mainly sunk-cost investment, competition cannot eliminate the resulting quasi rents, and a long-term contract may be needed to allocate them efficiently and restore the incentives for efficient investment.

### III. Investment Incentives with Symmetric Bargaining

Section II's results generally confirm the private provision of public goods intuition that short-term contracting yields inefficiently low levels of sunk-cost relationship-specific investment. This section tests the robustness of Section II's view of investment incentives by replacing its all-or-nothing offers model of bargaining with the symmetric Nash (1950, 1953) bargaining solution.

The results reinforce Section II's conclusions. Short-term contracting with symmetric bargaining interferes with consumption smoothing in essentially the same way as with all-or-nothing offers, and implies a similar role for the assignment of ownership of reversible capital. The analysis of a leading class of examples suggests that short-term contracting's tendency to yield inefficiently low levels of sunk-cost investment is almost as strong with symmetric bargaining as with all-or-nothing offers. In particular, the instances of overinvestment that occur in the all-or-nothing offers model might appear to stem from its failure fully to capture the private provision of public goods intuition, which assumes that *both* parties share in the surplus generated by investment. But overinvestment occurs in the same kinds of circumstances with symmetric bargaining, suggesting that the structure of second-best compromises between efficient investment and efficient consumption smoothing, and not how parties share the surplus, is the main determinant of the direction of investment bias.

Except for the all-or-nothing offers model of bargaining, this section's model maintains Section II's notation and assumptions. Bar-



gaining outcomes are described instead by the Nash bargaining solution, defined as follows. Let  $u$  and  $v$  denote  $U$ 's and  $V$ 's negotiated utilities,  $\underline{u}$  and  $\underline{v}$  their disagreement utilities, and  $S$  the utility-possibility set. When, as here,  $S$  contains a pair of utilities strictly preferred to disagreement by both parties, the Nash solution solves

$$(N) \max_{\substack{u \geq \underline{u} \\ v \geq \underline{v}}} (u - \underline{u})(v - \underline{v}) \quad \text{s.t. } (u, v) \in S.$$

The Nash solution generalizes the intuitively appealing surplus-sharing principle of equal-gains-from-disagreement to bargaining problems with nonlinear utility-possibility frontiers. Nash (1950) showed that it is the only function of the disagreement utilities and the utility-possibility set that is efficient, symmetric across bargainers, independent of increasing linear transformations of their von Neumann-Morgenstern utility functions, and independent of irrelevant alternatives in the sense that shrinking the utility-possibility set without removing the original bargaining solution does not alter the solution.

Nash (1953) postulated a bargaining model in which bargainers submit simultaneous "demands," with compatible demands yielding a binding contract and incompatible demands yielding disagreement. He then observed that any efficient outcome that both parties prefer to disagreement can be supported as an equilibrium in this game, and argued that the Nash solution can serve as what would now be called a focal point to help bargainers choose among these multiple equilibria. The solution concept used in this section can be viewed, following his argument, as a description of how parties choose among multiple bargaining equilibria in contract negotiations in which parties make simultaneous demands.

I now characterize the long- and short-term contracting outcomes with Nash bargaining when only sunk-cost investment is required for efficiency. Note first that, when  $U$  and  $V$  own nothing that has value outside their relationship, their undiscounted second-period disagreement utilities are  $u(x)$  and  $v(y)$ . (As before, they never find storage beneficial.) Thus, if they invest  $s$  in the first period

and are not bound by contract in the second period, they will agree to cooperate in production, with consumption shares determined by the Nash solution for the second-period bargaining environment, which solves<sup>9</sup>

$$(N') \max_{0 \leq d_2 \leq z(s)} [u(z(s) - d_2) - u(x)] \\ \times [v(d_2) - v(y)].$$

The constraints that the expressions in brackets must be nonnegative are omitted for clarity in  $(N')$ , and in  $(L')$  below. The solution of  $(N')$ , easily shown to be unique under my assumptions, is written  $d_2 = f[z(s)]$ ; its dependence on  $u(x)$  and  $v(y)$ , which are fixed throughout the analysis, is suppressed for clarity.<sup>10</sup>

Now reconsider Section II's arguments about  $U$ 's and  $V$ 's utilities if they fail to reach an agreement in the first period. If disagreement leads to autarky for two periods, it is clear that  $\underline{u} = (1 + a)u(x)$  and  $\underline{v} = (1 + b)v(y)$  for the reason given before. If disagreement enforces autarky for only one period,  $U$  and  $V$  rationally expect to reach agreement in the second period; this agreement will yield them  $u(z(0) - f[z(0)])$  and  $v(f[z(0)])$  under either long- or short-term contracting. Thus,  $\underline{u} = u(x) + au(z(0) - f[z(0)])$  and  $\underline{v} = v(y) + bv(f[z(0)])$  in this case. In what follows, I shall normalize  $\underline{u} = \underline{v} = 0$  in each case; this involves no loss of generality, because  $x$ ,  $y$ , and the assumption about how long an autarkic period follows a first-period disagreement remain fixed throughout.

<sup>9</sup>Second-period consumption can be determined in this way, by assuming that parties cooperate in production and applying the Nash solution, because efficiency requires cooperation and the Nash solution is therefore independent of the irrelevant alternative of not cooperating.

<sup>10</sup>The Nash solution traditionally gives parties' utilities. This is equivalent here to specifying their consumptions, which is more convenient for my purposes.

Given this normalization, the long-term contracting outcome solves

$$(L') \quad \max_{s, d_1, d_2 \geq 0} [u(z(0) - s - d_1) + a \\ \times u(z(s) - d_2)] [v(d_1) + bv(d_2)].$$

The short-term contracting outcome solves (L') with the added constraint that  $d_2 = f[z(s)]$ . (This problem need not have a unique solution in general, but it is well behaved whenever  $f[\cdot]$  is not too nonlinear.) These outcomes are characterized by first-order conditions with standard efficiency and surplus-sharing interpretations. The second-best efficiency condition for short-term contracting, for example, equates the marginal rate of transformation to a weighted average of parties' marginal rates of substitution, with the weights reflecting how the bargaining solution constrains them to share the fruits of their investment in the second period.

The main question of interest is whether Section II's view of the bias in sunk-cost investment under short-term contracting, based on the all-or-nothing offers model of bargaining, is robust to this section's change in bargaining solution. I am unable to offer a general answer to this question, but the following observations may shed some light on it.

First, it is evident that when parties are perfectly symmetric (that is, when  $u(\cdot) \equiv v(\cdot)$ ,  $a = b$ , and  $x = y$ ), efficient consumption smoothing is symmetric, and therefore coincides with the solutions to parties' symmetric bargaining problems under short-term contracting. Thus, the constraint short-term contracting imposes on parties' agreements is not binding, and long- and short-term contracting are equivalent even though parties have equal discount factors.<sup>11</sup>

<sup>11</sup>In Section II's model, parties with equal discount factors always invest too little, so overinvestment can occur only when their discount factors differ significantly. The result just described suggests that it may be possible to find examples in which symmetric bargaining yields overinvestment even when parties have equal

The striking difference between this result and the systematic downward investment bias in symmetric environments with all-or-nothing offers may make Section II's conclusion that underinvestment is the norm appear to be an artifact of its extreme assumption about surplus-sharing. Extensive efforts to prove this have convinced me that it is not true, without yielding a general argument for this conclusion. The conclusion is valid, however, for a wide class of tractable examples. Suppose that one party, say  $V$ , has linear preferences, so that  $v(d) \equiv d$ . Then it can be shown, by analyzing the first-order conditions that characterize the long- and short-term contracting outcomes, that large differences in parties' discount factors can make overinvestment a useful commitment device, but that parties who have similar discount factors always underinvest. This class of examples does not appear unrepresentative, but the analysis does not indicate whether its conclusion should be expected to generalize. This section's analysis therefore lends only limited further support to the private provision of public goods intuition.

The Nash bargaining solution responds to changes in parties' disagreement utilities in the natural way: the better a party can do on his own, the larger his share of the surplus in bargaining, other things equal. Thus, as in Section II's model, the assignment of ownership of reversible capital can be used as a substitute for contractual commitments about future compensation. Proposition 3's conclusion therefore remains valid with symmetric bargaining.

discount factors, by perturbing symmetric environments in which there is no investment bias. For, the levels of investment under long- and short-term contracting are, under my assumptions, differentiable functions of parameters that affect the environment differentially. Because they are determined by different systems of equations, they will in general have different derivatives with respect to such a parameter. Unfortunately, it can be shown that these derivatives are necessarily equal at parameter values that imply symmetry, so this approach does not reveal whether symmetric bargaining can yield overinvestment with equal discount factors.

#### IV. Conclusion

This section discusses the paper's relationship to the literature. At the most abstract level, the questions studied here have to do with the value of commitments in dynamic games. The usual arguments that such commitments have value (see, for example, Finn Kydland and Edward Prescott, 1977) are inconclusive here, because the ability to make complete short-term contracts gives parties much more control over the organization of their relationship than those arguments assume. Proposition 1 shows that this degree of control suffices for efficient organization of many relationships in which long-term commitments would have value if parties could not make complete short-term contracts.

The questions in contract theory posed here also parallel those considered by Oliver Hart (1975) in the theory of competitive general equilibrium. Hart studied pure-exchange economies with perfect information and perfect foresight, in which there are complete spot markets in each period, but no futures markets to allow long-term commitments. He provided robust examples in which sequences of competitive spot-market equilibria yield inefficient allocations, because competition on spot markets alone does not compel (or generally allow) rational agents to equate their marginal rates of substitution across time periods. Hart's examples suggest that long-term commitments might also be beneficial in contractual relationships. But individual contractual relationships are free of the market interactions that drive his results, and my analysis shows that long-term contractual commitments have value partly for different reasons.

Several recent papers study the uses of long-term contracts in dealing with the incentive problems caused by asymmetric information; see, for example Jean Tirole, 1986; Fudenberg, Holmstrom, and Milgrom, 1987, and the references given in those papers. A central theme of this literature is the value of "memory" in incentive contracts. Long-term contracts allow commitments about how information revealed now

will be used in the future that enables parties to deal more effectively with asymmetric-information incentive problems. When, as here, these are eliminated by assuming that parties have perfect or symmetric information, long-term contracts cannot serve this purpose.

The literature that deals specifically with investment incentives includes papers by Klein, Crawford, and Alchian, 1978; Paul Groot, 1984; Tirole, 1986; and Williamson, 1979, 1983. An important difference between these analyses and mine is that I allow parties to contract for the level of investment as well as other relevant variables.<sup>12</sup> The investment biases studied here therefore have some claim to be considered more basic.

Turning to still more closely related work, the private provision of public goods analogy may not always predict the direction of sunk-cost investment bias in Section II's model because the analogy assumes that current compensation for investment is impossible, whereas the two-period model allows compensation over a significant portion of the investment's life. This suggests that a stronger result might be available in a multi-period model, in which investment and consumption proceed continually and short-term contracting allows parties to make only insignificant commitments about compensation. My 1987 paper develops a continuous-time model, in which short-term contracting is instantaneous, to learn whether this is true, and to test the robustness of the present paper's results.

The analysis yields a strikingly simple explanation of how the duration of contracts influences sunk-cost investment. By restricting parties to a particular time pattern of compensation for any given investment plan, short-term contracting interferes with consumption smoothing within their relationship. This tends to make them collectively less tolerant of time variation in the relationship's output than they would be under long-term contracting. As in single-agent models, this reduces the optimal level of

<sup>12</sup> Tirole (1986) is a partial exception.

investment. Because the long-term contracting outcome is efficient, it follows that short-term contracting tends to yield too little sunk-cost investment for efficiency. The analysis indicates that the results obtained in this paper are not simply artifacts of its two-period framework, but shows that the incentives to invest in a relationship under short-term contracting are more complex than is apparent from the private provision of public goods analogy.

Another related paper, by Norman Clifford and myself (1987), studies the effect of short-term contracting on the exploitation of an exhaustible resource. It is shown there that the inability to make long-term contracts tends to cause inefficiently slow extraction, for reasons like those just discussed.

Finally, in Section IV of my 1985 paper, the ideas presented here were used to interpret the pattern of use of compulsory interest arbitration, which covers policemen and firefighters—who enjoy a great deal of short-run “bargaining power” in their relationships with their employers—much more often than other types of employees. My analysis suggests that an arbitrator can restore the incentives for efficient investment in such relationships simply by remembering the relationship-specific investments parties made in the past and standing ready in new-contract negotiations to impose a settlement that provides adequate compensation for them. This solution has the attractive feature that it does not require the arbitrator to impose a settlement unless parties cannot agree on one. Thus, unlike many suggested applications of the theory of incentives, it does not interfere with their “right to bargain,” which seems to be inalienable in practice.

#### REFERENCES

- Becker, Gary, “Investment in Human Capital: A Theoretical Analysis,” *Journal of Political Economy*, October 1962 Supplement, 70, 9–49.
- Clifford, Norman and Crawford, Vincent P., “Short-Term Contracting and Strategic Oil Reserves,” *Review of Economic Studies*, April 1987, 54, 311–23.
- Crawford, Vincent P., “Dynamic Bargaining: Long-Term Relationships Governed by Short-Term Contracts,” Discussion Paper 83-13, University of California-San Diego, May 1983.
- , “The Role of Arbitration and the Theory of Incentives,” in Alvin E. Roth, ed., *Game-Theoretic Models of Bargaining*, London and New York: Cambridge University Press, 1985, ch. 17.
- , “Long-Term Relationships Governed by Short-Term Contracts,” Working Paper No. 205, Industrial Relations Section, Princeton University, February 1986.
- , “Relationship-Specific Investment,” unpublished manuscript, University of California-San Diego, November 1987.
- Dow, Gregory K., “Internal Bargaining and Strategic Innovation in the Theory of the Firm,” *Journal of Economic Behavior and Organization*, September 1985, 6, 301–20.
- Fudenberg, Drew, Holmstrom, Bengt and Milgrom, Paul, “Short-Term Contracts and Long-Term Agency Relationships,” Working Paper No. 488, MIT, June 1987.
- Grout, Paul, “Investment and Wages in the Absence of Binding Contracts: A Nash Bargaining Approach,” *Econometrica*, March 1984, 52, 449–60.
- Hart, Oliver D., “On the Optimality of Equilibrium When the Market Structure Is Incomplete,” *Journal of Economic Theory*, December 1975, 11, 418–43.
- Klein, Benjamin, Crawford, Robert and Alchian, Armen, “Vertical Integration, Appropriable Rents, and the Competitive Contracting Process,” *Journal of Law and Economics*, October 1978, 21, 297–326.
- Kydland, Finn and Prescott, Edward, “Rules Rather than Discretion: The Inconsistency of Optimal Plans,” *Journal of Political Economy*, June 1977, 85, 473–92.
- Mas-Colell, Andreu, *The Theory of General Economic Equilibrium: A Differentiable Approach*, London and New York: Cambridge University Press, 1985.
- Nash, John, “The Bargaining Problem,” *Econometrica*, April 1950, 18, 155–62.
- , “Two-Person Cooperative Games,”

- Econometrica*, January 1953, 21, 128-40.
- Tirole, Jean, "Procurement and Renegotiation," *Journal of Political Economy*, April 1986, 94, 235-59.
- Williamson, Oliver, "Transaction-Cost Economics: The Governance of Contractual Relations," *Journal of Law and Economics*, October 1979, 22, 233-61.
- \_\_\_\_\_, "Credible Commitments: Using Hostages to Support Exchange," *American Economic Review*, September 1983, 73, 519-40.

# Experimental Tests of the Separation Theorem and the Capital Asset Pricing Model

By YORAM KROLL, HAIM LEVY, AND AMNON RAPOPORT\*

*A computer-controlled multistage portfolio selection task was devised to test experimentally basic assumptions underlying the separation theorem and the capital asset pricing model. Although most of the subjects diversified among the three risky assets, the introduction of a riskless asset did not have the effect predicted by the separation theorem, nor were the subjects affected by systematic changes in the variance-covariance matrix of the risky returns. However, performance improved as the reward was increased tenfold.*

The capital asset pricing model (CAPM), proposed by William Sharpe, 1964; John Lintner, 1965; and Jan Mossin, 1966, has been the predominant normative as well as positive theory for determining assets' returns and prices in financial markets. The CAPM has a wide range of implications for various financial issues including the allocation of risky assets, price predictions, performance indices, and market efficiency tests, as well as tools for corporate finance decisions such as capital structure, costs of financial sources, and prices of new issues. During the last decade numerous theoretical studies have examined and extended the CAPM. A most recent development is the arbitrage pricing model, proposed by Stephen Ross (1976), which constitutes a more general model that yields the CAPM as a special case.

Many empirical studies have been designed to test the CAPM, but all of them have come under the sharp criticism of Richard Roll (1977), who argued convincingly that the only way to assess the validity of the CAPM is by testing whether the

market portfolio is mean-variance (M-V) efficient. As Roll noted, this is an impossible task because the market portfolio ought to include all the available risky assets and, therefore, it can hardly be identified. Even if it could be identified, there is a technical problem: an M-V efficient set with so many assets cannot be derived.

An alternative approach is to test whether investors' behavior under well-controlled laboratory conditions is accountable for by the M-V model (see Harry Markowitz, 1952, 1959, and James Tobin, 1958), which underlies the CAPM. We describe below two experiments in which highly motivated subjects, who possess complete information about the parameters governing the portfolio selection task, are asked individually to allocate actual amounts of money in a series of portfolio selection tasks. Under these simplified but highly controlled experimental conditions, we examine whether the actual investment decisions conform to or approach with practice the ones predicted by the CAPM, or if other regularities and psychological principles underlie the behavior of our subjects.

In both experiments, each subject was first asked to distribute his or her investment capital in a series of 200 portfolio selection trials, each of which included three risky assets whose distributions of return were known and fixed over time. In 100 additional trials, a riskless asset with the same interest rate for borrowing and lending was introduced. The first experiment included

\*Kroll, The Hebrew University of Jerusalem and the Ruppin Institute, Israel; Levy, The Hebrew University of Jerusalem and the University of Florida at Gainesville; Rapoport, The University of North Carolina at Chapel Hill and the University of Haifa, Israel, respectively. We would like to thank Ariel Cohen for help in computer programming and Lyle V. Jones for helpful comments. Helpful comments from two anonymous referees are also gratefully acknowledged.

three between-subject conditions, which differed from one another in the variance-covariance matrix governing the risky returns. The second experiment replicated one of the three conditions of the first experiment with a considerably more highly paid group of subjects. The two experiments were designed to investigate four major issues: (1) the effects of the correlations between the risky assets on the investment portfolios; (2) the Separation Theorem, an essential ingredient of the CAPM, which states that under certain assumptions all investors who believe they face the same efficient frontier and riskless lending and borrowing rates, regardless of their attitude toward risk, ought to hold the same portfolio of risky assets and differ from one another only in the proportion of investment capital placed in the riskless asset; (3) the effects of magnitude of reward on the investors' behavior; and (4) the magnitude of individual differences.

Section I describes the two portfolio selection tasks employed in both experiments. The M-V opportunity set and the optimal investment decisions are presented in Section II. Section III describes and discusses the results of Experiment 1, and Section IV does the same for Experiment 2. Concluding remarks are stated briefly in Section V.

### I. Experimental Procedure and Task Description

Experiments 1 and 2 each included two different tasks, called Task A and Task B, which were conducted in three separate experimental sessions. Task A was administered in Session 1 and fully replicated in Session 2. Task B was conducted in Session 3 only. A brief description of the two tasks is presented below; for a complete description see the Appendix.

*Task A.* Task A consisted of ten identical and mutually independent computer-controlled portfolio selection problems (games) with a maximum of 10 trials each. These were relatively simple investment problems with no short sales, transaction costs, and income or capital gains tax. Each trial in each of the ten games allowed the subject to invest in one or more of three

risky alternatives (stocks) *A*, *B*, and *C* with normal return distributions  $N(3,3)$ ,  $N(5,6)$ , and  $N(7,12)$ , respectively. The three return distributions remained unaltered during the three experimental sessions. The subjects were fully informed about the parameters of the three return distributions and the correlations between the returns.

On the first trial of each of the ten problems in Task A, the subject was provided with an initial capital,  $x_1$ . The same value was used for each subject and each problem. The subject was then required to distribute  $x_1$  over the three risky assets. Prior to making the investment decision, the subject was allowed to observe and examine (by pressing an appropriate button on the computer keyboard) for no cost at all the previous returns of stocks *A*, *B*, and *C* in up to the last 60 trials. Following the investment decision on trial 1, the returns for stocks *A*, *B*, and *C* were sampled randomly from their respective distributions, the starting capital for trial 2,  $x_2$ , was calculated according to the realized returns, and the task proceeded to trial 2. All the ten trials in a problem had the same structure: a display of information about the problem number, trial number, number of units of each stock, capital invested in each stock, and total investment capital; an optional request for information about previous returns; a choice of portfolio (see the Appendix). At the end of the tenth trial of each problem, provided the subject was not bankrupt earlier ( $x_t \leq 0$ ,  $t = 1, \dots, 10$ ), the profit was calculated (i.e., the difference  $x_{11} - x_1$  was computed). The differences  $x_{11} - x_1$  were then summed over the ten problems, converted from the currency used in the experiment to money according to a prespecified and known rate, and paid to the subject.

*Task B.* Task B was identical to Task A except for important added features. In addition to investing in the risky assets, on each trial  $t$  the subject could either invest any part of his or her capital  $x_t$  in a riskless asset with a certain return of 2 percent per trial, or borrow at 2 percent for the purpose of as much as approximately four times the value of  $x_t$ . As in Task A, the computation of the new investment capital,  $x_{t+1}$ , was

TABLE 1—MAXIMUM NUMBER OF PORTFOLIO DECISIONS OVER SUBJECTS IN EXPERIMENTS 1 AND 2

Correlation Values <sup>a</sup>	Experiment 1			Maximum <sup>b</sup> Number of Decisions Over Games
	Task A: No Riskless Asset		Task B: With Riskless Asset	
	Games 1–10	Games 11–20	Games 21–30	
Group <sup>c</sup> 1 $\rho_{BC} = 0$	1000	1000	1000	3000
Group 2 $\rho_{BC} = 0.8$	1000	1000	1000	3000
Group 3 $\rho_{BC} = -0.8$	1000	1000	1000	3000
Correlation Value	Experiment 2			Maximum Number of Decisions Over Games
	Task A: No Riskless Asset		Task B: With Riskless Asset	
	Games 1–10	Games 11–20	Games 21–30	
Group 4 $\rho_{BC} = 0$	1200	1200	1200	3600

<sup>a</sup>For all four groups  $\rho_{AB} = \rho_{AC} = 0$ .

<sup>b</sup>There are 10 games in each session and a maximum of 10 trials per game for a maximum of 3000 portfolio decisions over the 30 subjects.

<sup>c</sup>There are 10, 10, 10, and 12 subjects in groups 1, 2, 3, and 4, respectively.

displayed on the CRT in front of the subject (see the Appendix).

The amount borrowed had to be returned with interest at the end of the same trial. However, if the subject borrowed money and invested the borrowed funds plus the investment capital in the risky assets, bankruptcy was possible. (Bankruptcy occurred when the total terminal value of stocks *A*, *B*, and *C* was smaller than the amount which had to be paid on the loan.) When bankruptcy occurred, the problem was terminated and the subject moved to the next game.

In order to investigate the effect of differing variance-covariance structures, the 30 subjects in Experiment 1 were divided randomly into three groups (1, 2, and 3) of 10 subjects each. The correlation between the returns on stocks *B* and *C*,  $\rho_{BC}$ , was set to be 0, 0.8 and  $-0.8$  for groups 1, 2, and 3, respectively. For each group,  $\rho_{AB} = \rho_{AC} = 0$ .

Conditions in Experiment 2 were identical to those for group 1 in Experiment 1 with two exceptions: the reward was increased tenfold, and the number of subjects was changed from 10 to 12.

Each of the 42 subjects in both experiments participated individually in three self-paced sessions, each of which lasted between 60 to 120 minutes. Two to seven days

elapsed between consecutive sessions. The mean payoff for all three sessions was \$18.91 in Experiment 1 and \$165 in Experiment 2.

The experimental design is summarized in Table 1.

**Experimental Procedure.** Tasks A and B were both administered to each subject individually by a Digital Equipment Corporation PDP-11 computer. The subject was seated in front of a computer terminal in a separate cubicle and was asked to read a set of instructions specifying that the purpose of the experiment was "to learn how people invest their capital between risky and riskless assets whose prices change over time" (see the Appendix).

The instructions explained the two tasks in detail, and then used several examples to demonstrate the effects of borrowing, lending, and dividing the investment capital among the risky assets in alternative ways (see the Appendix). In addition, the following points were emphasized:

1. Each session includes ten games of a maximum of ten trials each. Successive games are independent.

2. The three distributions of the risky returns are normal with means of 3, 5, and 7 percent, and standard deviations of 3, 6, and 12 percent for assets *A*, *B*, and *C*, respec-



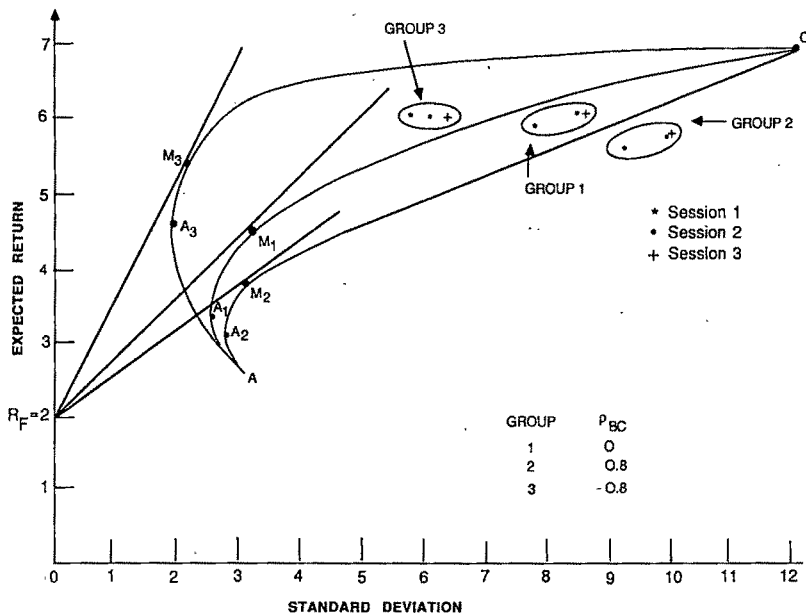


FIGURE 1. SINGLE-PERIOD EFFICIENT FRONTIERS FOR THE THREE EXPERIMENTAL CONDITIONS IN EXPERIMENT 1, AND MEANS AND STANDARD DEVIATIONS OF ACTUAL PORTFOLIOS BY GROUP AND SESSION

tively. The distributions remain unaltered for all the games.

3. Successive trials within games are independent. Namely, returns are drawn at random from the above normal distributions.

4. Subjects in group 1 were instructed that the return rates for the three risky assets  $A$ ,  $B$ , and  $C$  were statistically independent. Subjects in groups 2 and 3 were told that the return rate on asset  $A$  was independent of  $B$  and  $C$ , but that there was a positive correlation of 0.8 (for group 2) or a negative correlation of  $-0.8$  (for group 3) between the return rates of stocks  $B$  and  $C$ .

5. The total reward was contingent upon performance. The points gained during the 30 games in the three sessions were converted into money, which was paid to the subjects at the end of Session 3.

## II. The Optimal Investment Policy

### A. The Unleveraged Single-Period Opportunity Set

Figure 1 depicts the single-period efficient frontiers for groups 1, 2, and 3 in Experi-

ment 1. The curved line from  $A_1$  to  $C$  is the locus of the mean-standard deviation efficient unleveraged portfolios for group 1 ( $\rho_{BC} = 0$ ). Similarly, the curves from  $A_2$  to  $C$  and from  $A_3$  to  $C$  represent the efficient unleveraged frontiers for groups 2 ( $\rho_{BC} = 0.8$ ) and 3 ( $\rho_{BC} = -0.8$ ), respectively.

Each of the mean-variance efficient frontiers displayed in Figure 1 consists of all the portfolios that minimize the portfolio's variance for a given mean return. When short sales are prohibited (as is the case in our two experiments), quadratic programming is used to solve for the efficient frontier. Specifically, we solve for the following problem:

$$(1) \text{ Minimize } \underline{P}'\Sigma\underline{P}$$

subject to the constraints

$$(a) \quad \underline{P}'\underline{\mu} = E,$$

$$(b) \quad \underline{P}'\underline{1} = 1,$$

$$(c) \quad \underline{P} \geq 0,$$

where  $\Sigma$  is an  $n \times n$  variance-covariance matrix ( $n = 3$  in our two experiments),  $\underline{\mu}$  is an

$n \times 1$  vector of the mean returns of the  $n$  risky assets,  $\underline{P}$  is an  $n \times 1$  vector of the investment proportions, and  $E$  is the mean return of a given portfolio. The mean-variance frontier is generated by varying  $E$ , so that each point on the frontier corresponds to some vector of investment proportions.

Figure 1 illustrates the well-known fact that for a given standard deviation, the expected return from optimal diversification is higher as  $\rho_{BC}$  decreases, provided all the other parameters remain the same.

When riskless lending and borrowing at the same rate is allowed, we solve for the same problem, after replacing the constraints (a) and (b) in equation (1) by the constraint

$$(a') \quad \underline{P}'\underline{\mu} + (1 - \underline{P}'\underline{1})r = E,$$

where  $r$  denotes the riskless interest rate. Note that the sum of the proportions invested in the risky assets,  $\underline{P}'\underline{1}$ , is no longer equal to unity. The individual can invest more than 100 percent of his or her wealth in the risky assets (by borrowing) or less than 100 percent (by lending a portion of the wealth). Nevertheless, the total return from investing in the risky assets,  $\underline{P}'\underline{\mu}$ , plus (minus) the return from lending (borrowing) should be equal to the portfolio mean return,  $E$ .

For a given riskless interest rate, the investment proportions can be normalized such that  $P_i^* = P_i / \Sigma P_i$ . In this case,  $\Sigma P_i^* = 1$ . Hence, the investment proportions  $P_i^*$  correspond only to the proportions between risky stocks.

Recall that the same risky assets  $A$ ,  $B$ , and  $C$  were employed in Task B. Additionally, the subject had a choice between investing any part of his or her investment capital in a riskless asset with a safe return of 2 percent or borrowing at the same rate up to roughly four times the equity and then investing in the risky assets. If borrowing is not constrained (or, alternatively, the ceiling on borrowing is not binding) and the rates of borrowing and lending are the same, the efficient risky frontier is reduced to a single-optimal portfolio of risky stocks. This result is the core of the celebrated "Separation

Theorem." To explain this result, suppose that borrowing and lending are not allowed, as in Task A. Then the subjects in groups 1, 2, and 3, who are fully informed about the parameters of the investment task and attempt to maximize expected utility, will choose portfolios that fall on the lines  $A_1C$ ,  $A_2C$ , and  $A_3C$  in Figure 1, respectively. Different subjects within the same group are expected to choose different portfolios on the efficient frontier depending on the tradeoff they wish to attain between the expected return and "risk" (assumed to be measured by the standard deviation of the portfolio).<sup>1</sup> When the subjects can borrow and lend money freely at the same rate, the Separation Theorem states that all of them will distribute the portion of their capital invested in the risky assets in the same way. They may only differ from one another in the portion of capital invested in the risky assets. The Separation Theorem thus yields a strong invariance result, which holds for all risk-averse investors regardless of the shape of their utility functions.

Table 2 shows the optimal unleveraged portfolios for groups 1, 2, and 3 calculated by quadratic programming. Table 2 shows that as the correlation  $\rho_{BC}$  increases, the combined weight of stocks  $B$  and  $C$  in the optimal portfolios decreases, as does the expected return of the portfolio. Graphically, when borrowing is unconstrained, the optimal risky portfolio can be calculated by finding the tangent point to the risky efficient frontier of a straight line that intersects the vertical (expected return) axis at  $R_F = 2$  percent. Figure 1 portrays the tangent points  $M_1$ ,  $M_2$ , and  $M_3$  between the straight lines originating at  $R_F = 2$  percent and the efficient risky frontier for groups 1, 2, and 3, respectively.

To recapitulate, if borrowing and lending are unconstrained, the Separation Theorem implies the following testable prediction: In Task A different subjects may select different

<sup>1</sup>The Separation Theorem holds under the assumption that the subject only considers the three risky assets in the experiment. The effects of human capital and other sources of random income are ignored.

TABLE 2—OPTIMAL INVESTMENT PROPORTIONS IN THE THREE STOCKS FOR TASK B

$\rho_{BC}$	Optimal Proportions of Stock			Total	Portfolio Expected Return (Percent)	Portfolio Standard Deviation (Percent)
	A	B	C			
0	0.476	0.369	0.155	1.00	4.36	3.22
+0.8	0.580	0.400	0.020	1.00	3.88	3.12
-0.8	0.162	0.562	0.276	1.00	5.23	2.17

Note:  $\rho_{AB} = \rho_{AC} = 0$  in all three cases.

portfolios on the efficient frontier. When unconstrained borrowing and lending are allowed, we expect that risk-averse subjects will switch to a common, optimal, risky portfolio (Table 2), although they may still differ from one another in the proportion of their portfolio allocated to the riskfree asset. The subjects will do so if they only attend to the three risky assets in the experiment and ignore the effects of other risky assets (for example, human capital) that are not experimentally controlled. This prediction is experimentally testable in a within-subject design for each group separately even without measuring the utility functions of individual investors.

#### B. Constrained Borrowing and Multiperiod Considerations

The prediction above is based on the assumption that (a) borrowing is unconstrained and (b) that the task is single-staged. Both assumptions do not hold in Experiments 1 and 2: borrowing is constrained and a decision on trial  $t$  may change the investment capital and, hence, the investment decision on trial  $t + 1$ . However, both may be satisfied to a good approximation by most or all of our subjects. Thus, the ceiling on borrowing may not be perceived as binding, and subjects may perceive the multiperiod problem as a sequence of independent trials. Ample evidence has been accumulated to show that even in simple decision tasks subjects often simplify (or "frame") the task by using various cognitive heuristics (Daniel Kahneman and Amos Tversky, 1979; Kahneman, Paul Slovic, and Tversky, 1982). In multiperiod decision making, a useful and easily

implementable heuristic, which simplifies the task considerably, is to break the  $n$ -stage decision task into a sequence of  $n$ -independent single-stage tasks and perform each of them separately (Amnon Rapoport, 1984). Myopic behavior of this kind is in general suboptimal, leading to reduction in total profit. However, the losses it causes are often relatively small because optimal decision policies for multistage tasks are notoriously insensitive to deviations from them. On the other hand, myopic behavior reduces the computational time and mental effort required to solve the multistage task. Although myopic decision behavior is common in multistage decision tasks (for example, Rami Zwick and Rapoport, 1985), and may also be found in the present study, an examination of the optimal investment behavior under the multistage framework is warranted.

If borrowing is not limited, it can be shown that the optimal multiperiod portfolios are also combinations of the riskless asset and the *one-period*, optimal, unleveraged risky portfolios, which are presented in Table 2. In our experiments the decisions are multiperiod, but borrowing is constrained to about four times the capital at the beginning of each trial. We claim that even when borrowing is constrained, the inclusion of a riskless asset leads to optimal, unleveraged, risky portfolios which are either identical or close to the ones presented in Table 2.

If the amount borrowed is constrained, as is the case in the present study and most realistic investment situations, the efficient set includes more than a single-unleveraged portfolio. In this case, portfolios on the risky, unleveraged frontier with expected values that exceed the expected value of the

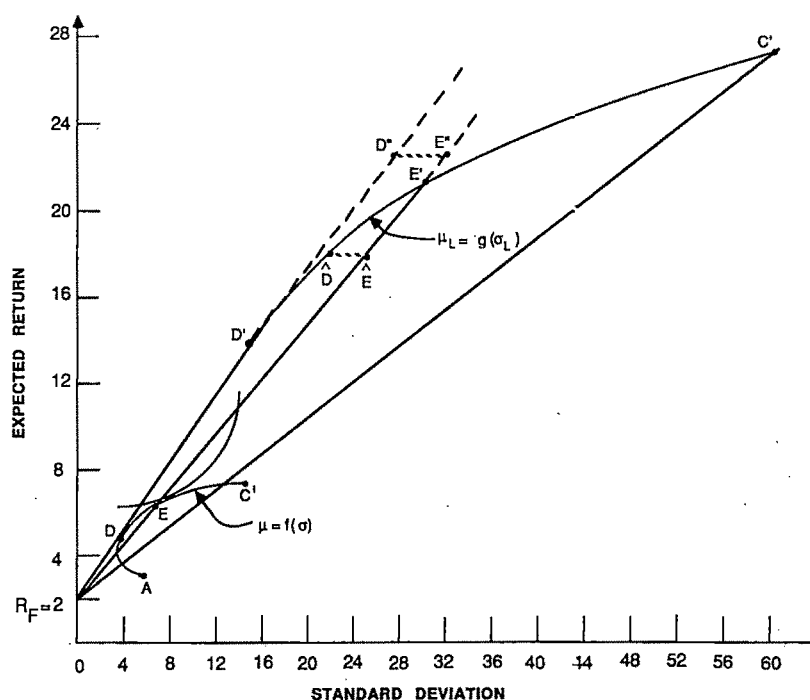


FIGURE 2. THE EFFICIENT FRONTIER FOR GROUP 1 WHEN BORROWING IS CONSTRAINED

optimal portfolio where borrowing is allowed are efficient. Figure 2 displays the efficient frontier for group 1 ( $\rho_{AB} = \rho_{AC} = \rho_{BC} = 0$ ) when the amount borrowed cannot exceed 3.9 times the amount of equity. The one-period efficient frontier is the straight line from  $R_F$  to  $D'$  and the curve that passes through the points  $D'$ ,  $E'$ , and  $C'$ . Note that points like  $D'$ ,  $E'$ , and  $C'$  are obtained by borrowing the maximum amount and investing all capital in the risky portfolios  $D$ ,  $E$ , and  $C$ , respectively. For example, investing the equity in stock  $C$  yields expected value and standard deviation of 7 and 12 percent, respectively. Borrowing the maximum amount and then investing the resulting sum ( $4.9x_t$ ) in stock  $C$  yields expected value and standard deviation of 26.5 and 58.8 percent, respectively (point  $C'$  in Figure 2).

An investor who chooses an optimal portfolio in Task A with expected return smaller than the expected value of portfolio  $D$  is expected in Task B to switch to point  $D$ . This investor may borrow or lend any frac-

tion of his or her capital at the riskfree rate of 2 percent and invest the rest in stocks according to the optimal risky portfolio  $D$ . It can be shown that an investor who chooses an optimal portfolio in Task A on the curve  $DC$  is likely in Task B to select an optimal portfolio closer to or on point  $D$ .<sup>2</sup> Thus, under constrained borrowing and multi-period considerations, not all investors will necessarily switch to a common optimal risky portfolio  $D$  as in the unconstrained case, but will, nevertheless, "move" along the efficient frontier toward the point  $D$ ; this result is experimentally testable.

### III. Experiment 1

*Subjects* Thirty male and female undergraduate students from the University of Haifa participated in Experiment 1. All the

<sup>2</sup>A proof is available from the authors upon request.

TABLE 3—MEANS AND STANDARD DEVIATIONS OF THE INVESTMENT PROPORTIONS OF THE STOCKS IN THE SELECTED UNLEVERAGED PORTFOLIOS

Session	Games		Group 1 ( $\rho_{BC} = 0$ )			Group 2 ( $\rho_{BC} = 0.8$ )			Group 3 ( $\rho_{BC} = -0.8$ )		
			A	B	C	A	B	C	A	B	C
1	1-10	<i>M</i> <sup>a</sup>	0.15	0.26	0.59	0.18	0.17	0.65	0.14	0.28	0.57
		<i>SD</i>	0.10	0.08	0.14	0.17	0.10	0.24	0.09	0.11	0.12
2	11-20	<i>M</i>	0.10	0.22	0.69	0.15	0.18	0.67	0.12	0.25	0.62
		<i>SD</i>	0.08	0.08	0.14	0.13	0.15	0.24	0.08	0.11	0.13
3	21-30	<i>M</i>	0.11	0.20	0.69	0.13	0.20	0.67	0.12	0.27	0.61
		<i>SD</i>	0.10	0.09	0.16	0.11	0.18	0.25	0.10	0.11	0.13

<sup>a</sup>*M* = Mean; *SD* = standard deviation. In all three groups  $\rho_{AB} = \rho_{AC} = 0$ .

subjects were volunteers who agreed to participate in a 3-session, computer-controlled portfolio selection experiment with monetary reward contingent on performance. Only subjects who satisfied the following two criteria were recruited. First, the subjects had completed at least a 1-year long course in statistics. Consequently, when the experiment commenced, the subjects were already familiar with the concepts of statistical independence, random sampling, normal distribution, and linear correlation.<sup>3</sup> Second, the subjects had participated in a similar, though considerably simpler, portfolio selection experiment conducted at the same laboratory. This earlier experiment included only a single-risky alternative and precluded borrowing, whereas the present experiment included three risky alternatives and allowed for both borrowing and lending. Participation in the earlier portfolio selection experiment familiarized the subjects with the computer console, the procedure for obtaining information about normally distributed returns, and trial-to-trial random changes in the return on risky assets.

*Game Effects.* Preliminary analyses showed no significant differences among the ten games in Session 1 in terms of the gain per game, mean investment proportions per game (computed over the ten trials in each game), and number of times per game that

information about past performance of the stocks was requested. Similarly, no game effects were found in Sessions 2 and 3. Consequently, statistics for individual subjects were computed on the basis of all the trials within each of the three sessions. Although the experimental design allowed for a maximum of 100 trials per subject and session, the number of trials was occasionally smaller due to bankruptcy. In Sessions 1 and 2 each of the 30 subjects completed all trials without going bankrupt. However, seven subjects went bankrupt during one or at most two problems in Session 3. As a result, the individual statistics presented and discussed below are based in almost all cases on 100 trials per session and in no case on fewer than 85 trials.

*Individual Portfolios.* The mean-investment proportions in stocks *A*, *B*, and *C* were computed separately for each subject over all the trials in each session. Hereafter, we shall refer to these mean results as the *individual portfolios*. Table 3 presents the means and standard deviations of the risky assets in the individual portfolios for each session separately.

The results displayed in Table 3 provide answers to two of the major questions raised above. First, there is no evidence for significant differences between the *mean* proportions of stocks of the three different groups. The differences in the correlation  $\rho_{BC}$  among the three groups do not seem to affect the mean-investment proportions. Second, there is no evidence for significant differences between the three sessions. Neither learning (session to session changes) nor the inclusion of the opportunity to lend and borrow mon-

<sup>3</sup>The statistical knowledge of the subjects was not assessed directly. We made no attempt to screen subjects on the basis of their understanding of basic concepts of statistics, or to discuss the concepts of mean, variance, correlation, normal distribution, and random sampling in the instructions.

ey seems to affect the mean-investment proportions.

An elaboration of the second finding is warranted. Recall that the present study set out to test the major prediction that risk-averse investors will switch in Session 3 to a common risky portfolio (assuming that borrowing and lending are unconstrained). In Section II we claimed that when the constraint on borrowing is binding, risk-averse investors would tend to move closer to the common, optimal, risky portfolio. However, inspection of Table 3 shows that, for each of the three experimental groups, the inclusion of riskless borrowing and lending did not decrease the between-subject standard deviations of the individual unlevered portfolios. For example, the standard deviations in group 1 for stocks *A*, *B*, and *C* in Session 2 are 0.08, 0.08, and 0.14, respectively, compared to 0.10, 0.09, and 0.16 in Session 3. No statistical tests are required to conclude that the inclusion of the riskless asset did not result in more homogeneous investment policy as predicted by the Separation Theorem. In comparison to the optimal risky portfolio shown in Table 2, the subjects invested on the average more heavily in stock *C* (69 vs. 16 percent, 67 vs. 2 percent, and 61 vs. 28 percent in groups 1, 2, and 3, respectively) and less in stocks *A* and *B*.

The tendency to invest more heavily in stock *C* is also apparent when the individual decisions on each trial are examined. The risky portfolio included only stock *C* in 3,572 (40 percent) of the total of 8,925 decisions in Sessions 1 through 3. On 4,838 (54 percent) trials, stock *C* was included in portfolios with two or more stocks. Only on 515 (6 percent) trials was stock *C* excluded from the portfolio. In comparison, stock *A* was excluded on 58 percent of all the trials in Sessions 1 through 3, and stock *B* was excluded on 44 percent of all the trials.

The means and standard deviations of the individual portfolio returns (computed over the three stocks) are portrayed in Figure 1 for each group and session separately. It is striking to note that the expected returns of the unleveraged individual portfolios are essentially the same for all three groups. Because all the three groups chose on the aver-

age the same portfolio (with almost identical expected return), the standard deviation of these unleveraged portfolios was determined by the correlation  $\rho_{BC}$ . Figure 1 shows clearly that the differences in standard deviation between the groups were substantial.

Borrowing occurred on 83.2 percent of all the trials in Session 3. Borrowing more than 95 percent of the maximum allowed occurred on 15.2 percent of all the trials. On the average, the subjects in groups 1, 2, and 3 borrowed 223, 204, and 221 percent of their equity, respectively. A one-way ANOVA conducted on the degree of leverage showed no significant differences among the three groups. Individual differences in the degree of leverage were substantial. They may be due to attitudes toward borrowing that individuals acquire during their lives. The relatively high level of borrowing that characterizes the average subject in our experiment may be explained by the relatively low level of risk that prevailed in the experiment; the only "collateral" against the loan was the profit in the game rather than the individual's own personal asset.

*Efficiency.* According to the Separation Theorem and the ensuing discussion, all the subjects in Session 3 ought to move toward a common portfolio, which maximizes the slope of the tradeoff line between standard deviation and expected return (i.e., the lines in Figure 1 that are tangent to the risky frontiers). This prediction can be tested by examining the observed slope of the tradeoff line, which is measured by  $s = (\mu - 2 \text{ percent})/\sigma$ , where  $\mu$  and  $\sigma$  are the expected return and standard deviation of the individual portfolio, respectively. When a riskless asset is allowed, the slope is expected to increase—a testable prediction.

A second measure of efficiency is the standardized "distance" between the selected and optimal portfolios. Define the measure

$$d' = (x_A - x_{A*})^2 + (x_B - x_{B*})^2 + (x_C - x_{C*})^2,$$

where  $x_{i*}$  is the optimal proportion of stock *i* in the risky portfolio, and  $x_i$  is the observed proportion ( $i = A, B, C$ ). Given the

TABLE 4—SLOPES ( $s$ ) AND DISTANCES ( $d$ ) BY SESSION AND GROUP

Session	Games	Group 1 ( $\rho_{BC}=0$ )		Group 2 ( $\rho_{BC}=0.8$ )		Group 3 ( $\rho_{BC}=-0.8$ )		
		$s$	$d$	$s$	$d$	$s$	$d$	
1	1-10	$M^b$	0.54	0.32	0.46	0.48	0.73	0.18
		$SD$	0.05	0.18	0.04	0.32	0.23	0.09
2	11-20	$M$	0.51	0.45	0.46	0.51	0.66	0.23
		$SD$	0.05	0.19	0.03	0.31	0.15	0.15
3	21-30	$M$	0.50	0.46	0.46	0.52	0.68	0.21
		$SD$	0.05	0.23	0.04	0.31	0.16	0.12
Optimal <sup>a</sup>			0.72	0.00	0.60	0.00	1.49	0.00

<sup>a</sup>Based on the optimal investment proportions in Table 2 for a riskless interest rate of 2 percent.

<sup>b</sup> $M$  = mean;  $SD$  = standard deviation.

risky portfolio ( $x_{A*}, x_{B*}, x_{C*}$ ), define the portfolio ( $x_{A**}, x_{B**}, x_{C**}$ ) as the one that maximizes the distance

$$d'' = (x_{A*} - x_{A**})^2 + (x_{B*} - x_{B**})^2 + (x_{C*} - x_{C**})^2.$$

The double-starred portfolio is the one "farthest" away from the optimal one. For example, if the optimal portfolio for group 1 is

$$x_{A*} = 47.6\text{percent} \quad x_{B*} = 36.9\text{percent}$$

$$x_{C*} = 15.5\text{ percent},$$

(see Table 1), then the one maximizing  $d''$  is given by

$$x_{A**} = x_{B**} = 0, \quad x_{C**} = 100\text{ percent}.$$

Finally, the standardized "distance" between a selected and an optimal portfolio is given by

$$(2) \quad d = d'/d'' \quad (0 \leq d \leq 1).$$

The slope  $s$  and standardized distance  $d$  were computed separately for each subject and each session. Table 4 presents the means and standard deviations of the measures  $s$  and  $d$  by group and session. There is no indication in Table 4 that the inclusion of a riskless asset in Session 3 affected either of these two measures. Neither the means nor

the standard deviations of either of these two measures changed significantly from Session 1 to 2 and from Session 2 to 3.

*Discussion.* If the CAPM is endowed with descriptive power, it is expected that (a) the population variance-covariance matrix governing the random rates of return will affect the portfolio, and (b) the change in portfolios from the condition when borrowing and lending are prohibited to the condition when borrowing and lending money at the same rate are allowed will be in the direction of a common, risky, optimal portfolio. We also hypothesized that (c) the observed individual portfolios will move in the direction of the efficient frontier as investors gain more experience in selecting portfolios and obtain more feedback about the consequences of their decisions. The results of Experiment 1 refute all three hypotheses.

There are several explanations for our findings that are neither exhaustive nor mutually exclusive. We discuss them without implying any order of importance.

1. It is tempting to argue that, although knowledgeable in elementary statistics, our subjects did not understand how to use the information about the correlation  $\rho_{BC}$  given to them in the written instructions. This argument loses much of its appeal when it is noted that the subjects requested and subsequently obtained information about past performance of the three stocks every five trials on the average in Session 1 and every six trials on the average in Session 2. Many of the subjects reported in a post-experimen-

tal interrogation that they scrutinized the information about the preceding rates of return very carefully in an attempt to discover "patterns" or "trends" in the rates of return. (For more evidence on this issue in a single-asset portfolio task, see Yoram Kroll, Haim Levy, and Rapoport, forthcoming 1988.) The subjects could, therefore, discern that the rates of return for stocks *B* and *C* were independent of each other in condition 1, positively and highly correlated in condition 2, and negatively and highly correlated in condition 3.

Nevertheless, despite the access to information, our results raise the question of whether investors accurately perceive correlations between pairs of stocks and take account of these correlations in selecting their portfolios, or simplify the task because of limitations on their cognitive capability and treat each stock in the portfolio independently of the others. The experimental literature about cognitive limitations and the use of heuristics in the solution of considerably simpler single-stage decision tasks under uncertainty would seem to support the latter alternative.

2. A second explanation that many economists may prefer dismisses the experimental findings as irrelevant because of the low stakes given to the subjects. It asserts that subjects are not sufficiently motivated to maximize gain because of the low stakes, or that M-V efficiency is not likely to obtain unless the stakes are much higher. Notwithstanding our impression that the subjects were highly motivated to maximize expected gain, there is no simple way to refute this argument because the stakes involved in portfolio selection experiments can never approach the ones involved in the selection of portfolios in the real world. However, in an attempt to partially meet this objection, we designed Experiment 2, to be presented below, replicating the conditions of group 1 in Experiment 1 with the only change that the stakes were increased tenfold.

3. It might be claimed that the heavy investment in stock *C* (the most profitable and risky stock) reflects an entirely different objective criterion, namely, maximization of the *geometric mean* of the returns. Maximiza-

tion of the geometric mean is equivalent to maximization of  $E \log(W)$ , where  $W$  is the terminal wealth. The geometric mean rule is justified if the utility function over wealth can be approximated by  $U(W) = \log(W)$ . In addition, this rule can be justified in the case of multiperiod horizon with repeated revisions (as is approximately the case in our experiment). In this case, it is claimed that a portfolio with a higher-geometric mean results in a higher probability to end with a higher-terminal wealth. Harry Markowitz (1976) discusses and analyzes the advantages and disadvantages of the geometric mean rule given by various studies.

In order to determine the investment strategy that maximizes the geometric mean of the returns, one has to solve the following problem:

$$(3) \quad \text{Maximize } E \left\{ \log \left( \sum_{i=1}^n P_i R_i \right) \right\}$$

subject to the constraints

$$P_i \geq 0, \quad i = 1, \dots, n$$

$$\sum_{i=1}^n P_i = 1,$$

where  $P_i$  denotes the proportion of wealth invested in the  $i$ th asset, and  $R_i$  is the return on asset  $i$  (assumed to be normal in our study). Because the explicit solution is quite complex, we employ an approximation which has the property that the smaller the return deviations around the mean, the better the approximation. This approximation will provide us with a crude idea about the change in the investment policy resulting from the introduction of a riskless asset when the subject's objective is to maximize the portfolio geometric mean return.

The expected geometric mean can be approximated by the Taylor expansion around the mean using the equation

$$(4) \quad E \log(1 + r) = \log(1 + E(r)) - 1/2 \left[ \sigma^2(r) / (1 + E(r))^2 \right].$$



TABLE 5—THE APPROXIMATED GEOMETRIC MEAN FOR SEVERAL POINTS ON THE UNLEVERAGED AND LEVERAGED FRONTIERS

Unleveraged Expected Return	Group 1 ( $\rho_{BC} = 0$ )			Group 2 ( $\rho_{BC} = 0.8$ )			Group 3 ( $\rho_{BC} = -0.8$ )		
	$\sigma$	GMU <sup>a</sup>	GML <sup>b</sup>	$\sigma$	GMU	GML	$\sigma$	GMU	GML
4.0	2.32	0.040	0.106	3.34	0.039	0.101	2.05	0.039	0.108
4.4	3.28	0.043	0.119	4.20	0.042	0.113	1.94	0.043	0.125
5.0	4.31	0.048	0.138	5.73	0.047	0.126	2.04	0.049	0.131
5.6	5.10	0.051	0.149	6.84	0.051	0.131	2.30	0.052	0.168
6.0	6.71	0.056	0.159	8.59	0.055	0.136	4.02	0.058	0.182
6.2	7.59	0.058	0.157	9.23	0.056	0.136	5.47	0.059	0.180
6.4	8.59	0.058	0.154	9.90	0.058	0.134	7.04	0.060	0.173
6.6	9.67	0.060	0.147	10.68	0.059	0.131	8.67	0.061	0.161
6.8	10.83	0.061	0.138	11.29	0.060	0.130	10.36	0.061	0.146
7.0	12.00	0.061	0.127	12.00	0.061	0.127	12.00	0.061	0.127

<sup>a</sup>The geometric mean of the unleveraged portfolio (GMU) is approximated by the equation  $E \log(1+r) = \log(1+E(r)) - \frac{1}{2}(\sigma^2(r)/(1+E(r))^2)$  where  $E(r)$  and  $\sigma(r)$  are the mean and standard deviation of the returns on the unleveraged frontier.

<sup>b</sup>The geometric mean of the leveraged portfolio (GML) is computed in the same way, but the means and standard deviations are those of the leveraged frontier.

Table 5 shows the geometric mean calculated according to the above equation for various points on the unleveraged and leveraged efficient frontiers. In the case where borrowing is not allowed, the highest expected geometric mean for all three groups is obtained by holding a portfolio with a mean return of about 7 percent. It is known that in general the maximum growth portfolio allows for diversification. Therefore, given the approximate nature of our solution (equation (4)), it is safe to assert that the portfolio maximizing the geometric mean in our study lies on the right-hand side of the efficient frontier with a relatively large proportion invested in asset C. However, when constrained borrowing is permitted, the maximum geometric mean for all the three experimental groups is a point on the efficient set located to the left of the previous solution, namely, a portfolio with a smaller proportion invested in stock C and higher proportions in stocks A and B.<sup>4</sup>

<sup>4</sup>Two remarks concerning the approximation in equation (4) are warranted. First, regardless of the magnitude of error introduced by the approximation, there is no reason to believe that for all three groups there are larger errors associated with Task A than Task

B. In contrast to the implications of the geometric mean rule, when borrowing was allowed the subjects did not reduce their investment in stock C and, in fact, hardly changed their investment diversification in the risky assets at all (see Table 3). This result refutes the hypothesis that the maximization of the geometric mean rule governed the portfolio decisions of our investors.

4. In the previous analysis we examined the impact of correlation and riskfree borrowing and lending on group behavior. We investigated the *mean*-investment proportions across subjects in each group and largely ignored the analysis of individual differences. It is possible that the impact of

B. Second, note that when borrowing and lending are unconstrained, the optimal investment strategy in the risky assets for all risk-averse subjects, and in particular for the ones with a logarithmic utility function, is presented in Table 2. As shown in this table, the proportion invested in security C is less than 0.28. It is obvious from Tables 2 and 5 that the maximum growth portfolio with no riskless asset has a higher proportion invested in security C than the one with a riskless asset. In the latter case, maximum growth is attained primarily by having more leverage rather than investing more heavily in security C.

correlation and riskfree opportunity to borrow and lend money cannot be detected in mean results, but that there exists "experts" whose investment behavior can be accounted for, at least to a first-order approximation, by portfolio theory.

As stated above, the subjects in our experiment could not mimic one another. However, "experts" in the marketplace are often mimicked by others. Therefore, if we identify such "experts" in our experiment, they may correspond to the ones who are actually mimicked in the marketplace.

Indeed, we found that four of the ten subjects in each of the three experimental groups performed markedly better than the remaining six subjects. For example, the mean-standard distance  $d$  between observed and predicted portfolios summed over the "experts" in each group and all three sessions was 0.377, 0.209, and 0.187 for groups 1, 2, and 3, respectively, compared to 0.539, 0.724, and 0.216 for the "nonexperts" in these three groups. Lacking independent criteria in this study to differentiate between "expert" and "nonexpert" investors, we do not pursue the analysis of the difference in performance between these two subgroups. In future studies we plan to establish such criteria and, by allowing free flow of information between subjects, examine individual differences more systematically and check whether nonexpert subjects mimic the behavior of expert subjects.

5. In addition to considerations based on the parameters of the task, the subjects might have searched for patterns in the returns over time. Searching for patterns and trends may reflect difficulties in comprehending the concept of random sampling, or may be generated by preconceived ideas about the behavior of stocks. Recall that the subjects repeatedly requested information about previous rates of return, which was completely useless given knowledge of the distributions of the returns. It is possible that the repeated requests for information about previous returns were intended not only to learn or verify the correlation  $\rho_{BC}$  but, more importantly, to gather data for generating, testing, and subsequently discarding hypotheses about patterns or trends in the returns over

time. Systematic studies of this possibility are warranted.

#### IV. Experiment 2

*Subjects.* Twelve male and female undergraduate students from the University of Haifa, hereafter referred to as group 4, participated in Experiment 2. All the subjects volunteered to participate in a four-session, computer-controlled, portfolio selection experiment with monetary reward contingent on their performance. As in Experiment 1, only subjects who had completed at least a 1-year long course in elementary statistics were recruited. (See fn. 3.) Unlike group 1 in Experiment 1, the subjects in group 4 had not participated in any portfolio selection experiment and therefore were not familiar with the portfolio selection task and the operation of the computer console. Consequently, Experiment 2 consisted of four rather than three self-paced sessions. Session 0 was for practice only; the subjects were encouraged to vary their investment decisions from trial to trial and were given a flat rate of \$3 for their participation in this session. Sessions 1 through 3 replicated the conditions of group 1 ( $\rho_{BC} = 0$ ) in Experiment 1. The mean payoff per subject for Sessions 1 through 3 was \$165, approximately \$44 per hour. This payoff was approximately 30 times higher than the student hourly rate that prevailed in Israel when the experiment was being conducted.

*Procedure.* Except for the higher stakes, the experimental procedure was identical to that of group 1.

*Results.* Preliminary analyses showed again no differences among the ten problems in each of the three sessions. Consequently, the portfolios of each subject were averaged over the 100 trials in each session to yield stable estimates. Table 6 presents the group means and standard deviations (in parentheses) of the individual portfolios, the degree of leverage (for Session 3 only), the slopes of the tradeoff line between risky and riskless opportunities, and the standardized distance from the efficient frontier. The same results for group 1 of Experiment 1 are presented for comparison.

TABLE 6—COMPARISON BETWEEN THE PORTFOLIOS OF GROUP 1 (EXPERIMENT 1) AND GROUP 4 (EXPERIMENT 2)

Session	Group <sup>a</sup>	Portfolio			Degree of Leverage	Slope (s)	Distance from Optimal Portfolio (d')
		A	B	C			
1	1	0.15	0.26	0.59	—	0.54	0.32
		(0.10)	(0.08)	(0.14)		(0.05)	(0.18)
	4	0.23	0.23	0.54	—	0.60	0.29
		(0.18)	(0.09)	(0.18)		(0.07)	(0.23)
2	1	0.10	0.22	0.69	—	0.51	0.45
		(0.08)	(0.08)	(0.14)		(0.05)	(0.19)
	4	0.15	0.23	0.62	—	0.59	0.34
		(0.10)	(0.10)	(0.14)		(0.09)	(0.18)
3	1	0.11	0.20	0.69	2.23	0.50	0.46
		(0.10)	(0.09)	(0.16)	(1.12)	(0.05)	(0.23)
	4	0.17	0.24	0.59	2.75	0.61	0.27
		(0.09)	(0.09)	(0.16)	(0.48)	(0.11)	(0.18)

<sup>a</sup>The reward in group 4 is ten times the reward in group 1. In both groups,  $\rho_{AB} = \rho_{AC} = \rho_{BC} = 0$ .

Groups 1 and 4 were first compared to each other in terms of the proportion of capital invested in stocks A and C. Table 6 shows that for all sessions the subjects in group 4 selected less risky portfolios on the average than group 1 with a lower proportion of capital invested in stock C. On the average, across all three sessions, subjects in group 4 invested 18.1 percent of their capital in stock A and 58.0 percent in stock C compared to 11.8 and 65.4 percent for group 1, respectively. A two-way group by session ANOVA yielded a significant group effect but neither a session effect nor a group by session interaction effect.

Table 6 shows that the mean degree of leverage in Session 3 was slightly higher for group 4 than group 1. The difference between the two groups is not significant, however. One might expect that due to the higher amount of money involved, subjects in group 4 would borrow less than subjects in group 1. On the other hand, recall that group 4 selected on the average a less risky unleveraged portfolio than group 1. Therefore, the almost identical mean degree of leverage for both groups implies that the leveraged portfolios of group 4 had, indeed, smaller variance than those of group 1.

Table 6 also shows that the mean slope of the line between risky and riskless opportunities was higher for group 4 than for group 1 in each of the three sessions. A two-way group by session ANOVA conducted on the slopes of the individual portfolios resulted in a significant difference between the two groups ( $p < 0.05$ ). Neither the session effect nor the session by group effect were significant. The significant group effect means that, on the average, subjects in group 4 diversified more efficiently than subjects of group 1. Similar conclusions emerge from a comparison of the two groups to each other in terms of the standardized distance from the efficient frontier.

Additional data (not reported in Table 6) show that the subjects of group 4 diversified more often than the subjects of group 1. Thus, the subjects of group 4 did not diversify at all and invested all of their capital in one stock on 27.8 percent of all the trials compared to 43.3 percent of all the trials of group 1. Similarly, the subjects of group 4 allocated their capital among the three risky assets on 57.8 percent of all the trials compared to 37.9 percent of all the trials of group 1. A  $2 \times 3$  group by session ANOVA conducted on the percentage of trials in

which capital was allocated among all three risky assets yielded a significant group effect; the effects due to session and to group by session interaction were not significant.

Another dependent variable that differentiates between the two groups is the number of times in a session that information about previous rates of return was requested. As noted above, given that subjects had complete information about the three distributions of return, this information was irrelevant. Although it was costfree, requesting it actually delayed the experiment and consequently decreased the gain per unit of time played. On the average, the subjects of group 4 requested this information three times as often as the subjects of group 1. This highly significant and very substantial difference between the two groups probably reflects more caution and care on the part of the subjects of group 4, a stronger tendency to search for and discover "trends" in the random rates of return, or both.

*Discussion.* The stakes in Experiment 2 were raised to a level seldom encountered in psychological investigations of decision behavior. Realizing that their potentially high earnings depended on their performance, the highly paid and highly motivated subjects of Experiment 2 reacted to the experimental manipulation by examining more frequently the past performance of the stocks, by diversifying more often their investment equity over the three risky assets, and by selecting less risky portfolios. However, because they selected on the average a higher degree of leverage, the subjects of group 4 ended up with more efficient individual portfolios in Session 3.

The significant differences reported above between groups 1 and 4 clearly indicate that the actual stakes involved in the experiment can affect the decision behavior of the subjects. More importantly, the direction of the effect is toward greater efficiency. If substantiated, this finding qualifies the findings of psychological studies of decision behavior which rely exclusively on questionnaire data or, when paying subjects, use very low stakes (for example Kahneman, Slovic, and Tversky, 1982; Kahneman and Tversky, 1979).

The major finding bearing on classical portfolio theory is that once again the Separation Theorem failed. Even with considerably higher stakes, the inclusion of a riskless asset in Session 3 did not render the subjects more homogeneous in terms of their risky portfolios.

## V. Concluding Remarks

With regard to the capital asset pricing model, the two experiments yielded both positive and negative results.

1) As predicted by the CAPM, in most cases the subjects diversified their investment capital among the three risky assets. However, on the average the subjects invested considerably more than predicted in the riskiest asset.

2) A tenfold increase in the amount of capital significantly improved the subjects' performance. As predicted by the CAPM, the subjects in Experiment 2 selected less risky portfolios and diversified their capital among the three risky assets more often than the members of group 1 in Experiment 1. This finding casts some doubt on the validity of the results of many experiments on decision making which involve trivial amounts of money or no money at all.

The next result might be interpreted as positive by practitioners in the marketplace and as negative by academicians who accept the random-walk hypothesis.

3) Over hundreds of trials, the subjects repeatedly requested information on past returns of the risky assets, even though they had complete information about the parameters of the distributions of return and basic understanding of the nature of random sampling. Although the information on past returns was costless, it slowed down the experiment and consequently reduced the mean payoff per unit time. This behavior may indicate mistrust in the instructions or misunderstanding of random sampling. More likely, it may reflect a deep-rooted tendency by investors to search for patterns even though they are informed that the "random-walk" hypothesis prevails. This finding may account for the flourishing industry of chartists and

technicians in the marketplace, despite the repeated claims that the behavior of stock prices may be described as a random walk.

On the negative side, we found the following results.

4) The population variance-covariance matrix that governs the risky assets had no significant effect on investment behavior.

5) The introduction of a riskless asset did not enhance homogeneity in investment behavior, in contradiction to the Separation Theorem.

These last two findings cast serious doubt on the validity of the Separation Theorem as well as of the CAPM. This conclusion would probably be even stronger if the experiments had been conducted with a larger number of risky assets. Subjects who have difficulties in handling a single correlation between two risky assets would not be expected to handle several such correlations which occur simultaneously.

Although the findings above pertain to laboratory microeconomies, that are necessarily very simple relative to those that have evolved naturally, there exist field data that tend to support our results. Marshall Blume, Jean Crockett, and Irwin Friend (1974) conducted an extensive survey showing that investors do not diversify according to the portfolio selection model. Indeed, a large proportion of the subjects ignore altogether the correlations between risky assets and invest only in a single security. The survey included 17,056 individual income tax forms and showed that 34.1 percent held only one stock, 50 percent held no more than two stocks, and only 10.7 percent of the investors held more than ten stocks.

Turning back to our study and attempting to defend portfolio theory, one could rightfully claim that our two experiments do not replicate precisely the conditions prevailing in the market. It might be argued that, although the experimental conditions that we used are ideal for testing the diversification principle, the experiments differ from actual market conditions in three major respects. In our two experiments the investment decisions are made individually, are not disclosed to the remaining subjects, and do not

affect the distributions of return on subsequent trials. Thus, we made no attempt in the present study to construct an experimental environment that closely resembles a multitrader market. Second, in actual trading in stocks and bonds the investment decisions of each participant typically become known to the other participants at some later stage. Quite possibly in actual trading markets there exists a small group of sophisticated investors whose investment decisions are mimicked or followed by the remaining, less sophisticated investors. This type of behavior is not possible in our study. Third, in actual trading markets, unsuccessful investors may go bankrupt and lose their own money, whereas in experimental games subjects may only lose money given to them by the experimenter.

Rather than perceiving these discrepancies as major flaws of our experimental design, we regard them as sources of ideas for future experimentation. Because the validity of the CAPM cannot be tested by field studies, systematic laboratory experiments should be conducted and more effort should be devoted to improving them. In particular, laboratory experiments designed to test the CAPM should manipulate experimentally (a) the amount of investment capital, (b) the interest rate for borrowing and lending, (c) the population variance-covariance matrix governing the behavior of the risky assets, (d) the magnitude of effect that the decisions of each investor have on subsequent returns, (e) the information that each investor has about the investment decisions and reputation of other investors, and (f) the financial knowledge and sophistication of the subjects. It would be useful and instructive to conduct such experiments with individuals who have acquired much experience in selecting portfolios. Although access to such a population of investors is difficult, it is by no means impossible.

#### APPENDIX: A DETAILED DESCRIPTION OF THE TWO INVESTMENT TASKS

Each subject was given a set of instructions that motivated the two investment tasks, explained the tasks in detail, illustrated them by examples, and explained

TABLE A1—AN EXAMPLE OF THE CRT DISPLAY FOR TASK A

Game 5				
Trial Number 2				
Name of Stock	Value of Stock <sup>a</sup>	Change <sup>b</sup>	Number of Units <sup>c</sup>	Value of Investment <sup>d</sup>
A	100.00	2.4	20.00	2000.00
B	80.00	-1.3	20.00	1600.00
C	50.00	7.4	10.00	500.00
Total Capital:				4100.00
Information? (Y/N)				
Investment:	A = 800	B = 1000	C = 2300	
No. of Units:	A = 8.00	B = 12.50	C = 46.00	
Are you satisfied with your decision? (Y/N)				

Game Number 5				
Trial Number 3				
Name of Stock	Value of Stock	Change	Number of Units	Value of Investment
A	101.80	1.8	8.00	814.40
B	84.88	6.1	12.50	1061.00
C	52.40	4.8	46.00	2410.40
Total Capital				4285.80
Information? (Y/N)				

<sup>a</sup> $v_{k,t}$  is the value of each unit of asset  $k$  on trial  $t$ .

<sup>b</sup> $R_{k,t}$  is the rate of return on one dollar invested in security  $k$  on trial  $t$  defined as

$$R_{k,t} = (v_{k,t} - v_{k,t-1}) / v_{k,t-1}.$$

<sup>c</sup> $n_{k,t}$  is the number of units of asset  $k$  held on trial  $t$ .

<sup>d</sup> $x_{k,t}$  is the amount of capital invested in asset  $k$  on trial  $t$  ( $x_{k,t} = n_{k,t} \cdot v_{k,t}$ ).

how to use the computer terminal. Rather than translating the instructions from Hebrew, we present below an abbreviated description of the two tasks.

The subjects were first instructed that the purpose of the experiment was to study how ordinary people invest their money in stocks and bonds whose prices change over time. They were further told that the experiment would involve three stocks and one bond (a riskless asset), and that the only information they would receive concerned the distributions of rate of return and past behavior of the three stocks.

Each subject was then told that he or she was about to participate in ten independent problems. Each problem was divided into ten trials. The structure of a trial differed from Task A to Task B.

**Task A.** Each trial in each game commenced by displaying on a CRT in front of the subject the following information:

$g$ —the game number ( $g = 1, \dots, 10$ );

$t$ —the trial number within game ( $t = 1, \dots, 10$ );

$v_{k,t}$ —the value of each unit of asset  $k$  on trial  $t$  ( $k = A, B, C$ );

$R_{k,t}$ —the rate of return on one dollar invested in security  $k$  on trial  $t$  defined as  $R_{k,t} = (v_{k,t} - v_{k,t-1}) / v_{k,t-1}$ ;

$n_{k,t}$ —the number of units of asset  $k$  held on trial  $t$ ;

$x_{k,t}$ —the amount of capital invested on trial  $t$  ( $x_{k,t} = n_{k,t} \cdot v_{k,t}$ );

$x_t$ —the total investment capital on trial  $t$  ( $x_t = \sum_k x_{k,t}$ ).

Table A1 presents an example of the display. On trial 2 ( $t = 2$ ) of game 5 ( $g = 5$ )  $v_{A,2} = 100$ ,  $v_{B,2} = 80$ , and  $v_{C,2} = 50$ . The subject possesses 20, 20, and 10 units of stocks  $A$ ,  $B$ , and  $C$ , respectively. Consequently, the investment in the three assets is  $x_{A,2} = 2000$ ,  $x_{B,2} = 1600$ , and  $x_{C,2} = 500$ . The total investment capital is  $x_2 = 2000 + 1600 + 500 = 4100$ . Table A1 shows that asset  $A$  went up by 2.4 percent from trial 1 to 2, asset  $B$  went down by 1.3 percent, and asset  $C$  went up by 7.4 percent.

After the information on asset values and investment capital was displayed on the CRT, the subject was provided an opportunity to acquire costfree information about the past performance of the three stocks. When the key "Y" (for "Yes") was pressed in response to the question "information?" (see Table A1), the values of  $R_{k,t}$ ,  $R_{k,t-1}$ , ...,  $R_{k,t-14}$  were displayed on the CRT for  $k = A, B, C$ . Pressing "Y" for a second time revealed the values of  $R_{k,t-15}$ ,  $R_{k,t-16}$ , ...,  $R_{k,t-29}$ . Altogether, by pressing the key "Y" up to a maximum of four times, the subject could inspect the realizations of the random rates of returns  $R_A$ ,  $R_B$ , and  $R_C$  in the 60 preceding trials for as much time as he or she wanted.

The subject was allowed to copy the information about past performance of the three stocks or summarize it on paper in any way he or she wished. The

return rates were randomly drawn from a multinormal distribution with a variance-covariance matrix that was known to the subject. For any subject looking for sequential dependencies between trials (which, of course, did not exist) or entertaining hypotheses about "patterns" or "trends," the return rates  $R_{k,t}$ ,  $R_{k,t-1}$ , ..., were the only source of information. Pressing the key "N" (for "No") in response to the question "information?" reactivated the original screen and caused the subject to proceed to the next question.

Following the information acquisition phase, the subject was required to allocate his or her total equity investment  $x_t$  among the three risky assets in any way he or she wished. The subject did so by typing the amounts invested in assets  $A$  and  $B$ . These two decisions are denoted by  $d_{A,t}$  and  $d_{B,t}$ , respectively. The computer then checked whether  $d_{A,t} + d_{B,t} \leq x_t$ . If no, an error message was printed, and the subject was asked to repeat his or her investment decisions. If yes, the computer set  $d_{C,t} = x_t - (d_{A,t} + d_{B,t})$  and used the prices  $v_{k,t}$  to purchase the appropriate number of units of stock  $k$  for trial  $t+1$  ( $n_{k,t+1} = d_{k,t}/v_{k,t}$ ). Then the values of the three stocks were updated, and the computer advanced to the next trial.

Table A1 illustrates the subjects' investment decisions and their effect. In the example in Table A1, the subject's decisions are  $d_{A,2} = 800$ ,  $d_{B,2} = 1000$ , and  $d_{C,2} = 2300$ . Consequently,  $n_{A,3} = 8.00$ ,  $n_{B,3} = 12.50$ , and  $n_{C,3} = 46.00$ . As shown in Table A1, following the allocation of his or her investment capital, the subject was provided an opportunity to change the investment decision. If, for some reason, the subject was dissatisfied with his or her decisions before they were implemented, he or she could press the key "N" in response to the question "Are you satisfied with your decision?" and restart the investment phase. Pressing "Y" caused the subject to move to the next investment period (trial).

Continuing the example, Table A1 shows that in trial 3  $R_{A,3} = 1.8$ ,  $R_{B,3} = 6.1$ , and  $R_{C,3} = 4.8$ . As a consequence of the subject's decisions on trial 2, his or her investment capital increased in this example from 4100 on trial 2 to 4285.80 (a 4.53 percent increase) on trial 3.

Task B also consisted of 10 independent and identical investment games with a maximum of 10 trials each. Each game consisted of the same three risky assets  $A$ ,  $B$ , and  $C$  with the respective normal return distributions  $N(3,3)$ ,  $N(5,6)$ , and  $N(7,12)$  as before. In addition, borrowing up to roughly four times the amount of investment capital ( $x_t$ ) and lending was allowed.

**Task B.** Each trial  $t$  in Task B started by displaying to the subject on the CRT the values of  $g$ ,  $t$ ,  $v_{k,t}$ ,  $R_{k,t}$ ,  $n_{k,t}$ ,  $x_{k,t}$ , and  $x_t$ . In addition, it displayed the interest rate,  $R_F$ , for borrowing and lending.  $R_F$  was fixed at 2 percent for all games and all trials. Lending money was accomplished by investing capital in bonds with the number of bonds purchased (or sold) on trial  $t$  denoted by  $n_{b,t}$  and the value of each bond denoted by  $v_{b,t}$ . Thus, the total equity investment in Task B is given by

$$x_t = \sum_k x_{k,t} + x_{b,t},$$

where  $x_{k,t}$  is as defined before and  $x_{b,t} = n_{b,t} \cdot v_{b,t}$ .

Table A2 illustrates the display. The values of  $g$ ,  $t$ ,  $v_{k,t}$ ,  $R_{k,t}$ , and  $x_{k,t}$  are as in Table A1. In addition, there are  $n_{b,t} = 9$  units of bond in this example at a value of  $v_{b,t} = 100$  per unit. The total equity investment is  $x_2 = 4100 + 900 = 5000$ .

The information acquisition phase in Task B was identical to that in Task A. Following this phase, the subject was asked whether he or she wanted to borrow money. Denote the amount borrowed on trial  $t$  by  $b_t$ . To prevent unlimited borrowing, and in accordance with real-life portfolio selection situations where borrowing is always restricted,  $b_t$  was restricted to 3.9 times the value of  $x_t$ .

If the subject answered positively to the question "Do you wish to borrow?" the computer responded by asking "How much?" The subject answered by typing the amount  $b_t$  and then allocating the investment capital  $x_t$  among the three risky alternatives. Because the interest rate for borrowing and lending was identical, a decision to borrow precluded the option of investing in bonds (lending).

In the example in Table A2, the subject borrowed 9000 at 2 percent. Following that, the subject invested 6000, 3200, and 4800 in the risky assets  $A$ ,  $B$ , and  $C$ , respectively. Consequently,  $n_{A,2} = 60$ ,  $n_{B,2} = 40$ , and  $n_{C,2} = 96$ . As in Task A, after investing his or her capital, the subject could alter the allocation by responding negatively to the question "Are you satisfied with your decision?" or answer positively and move to the next trial.

Continuing this example, Table A2 shows that the return rates on trial 3 were  $R_{A,3} = 1.6$ ,  $R_{B,3} = 9.4$ , and  $R_{C,3} = -11.2$ . The amount borrowed plus the interest rate ( $b_t(1 + R_F)$ ) were deducted before the subject's next decision. Consequently, in the example in Table A2 the subject's capital after returning the loan decreased from  $x_2 = 5000$  on trial 2 to  $x_3 = 4679.20$  on trial 3 ( $13859.20 - 9000 \times 1.02 = 4679.20$ ).

If the subject answered negatively to the question "Do you wish to borrow?" he or she was required to allocate  $x_t$  in any way he or she wanted among the three risky and one riskfree asset. The experiment proceeded then as in Task A.

**Additional Information.** Following the detailed description of the two tasks, the subject was given another page of written instructions concerning the stocks. He or she was encouraged to read them carefully, and ask the experimenter for clarification, because the instructions "bear directly on your profit."

1. "There are 10 problems in each session. Each problem consists of 10 trials.

2. The changes in the price of stock  $A$  will be sampled randomly from a normal distribution that *will remain fixed throughout the experiment*. The distribution has a mean of 3 percent and standard deviation of 3 percent (approximately 95 percent of the changes will fall between -3 and 9 percent).

3. The changes in the price of stock  $B$  will be sampled randomly from a normal distribution that *will remain fixed throughout the experiment*. The distribution has a mean of 5 percent and standard deviation of 6 percent (approximately 95 percent of the changes will fall between -7 and 17 percent).

TABLE A2—AN EXAMPLE OF THE CRT DISPLAY FOR TASK B

Name of Stock	Game Number 5 Trial Number 2		Number of Units	Value of Investment
	Value of Stock	Change		
A	100.00	2.4	20	2000.00
B	80.00	-1.3	20	1600.00
C	50.00	7.4	10	500.00
Bonds	100.00	2.0	9	900.00
Total Capital:				5000.00
Information? (Y/N)				
How much borrowing?	9000			
Investment:	A = 6000		B = 3200	C = 4800
No. of Units:	A = 60		B = 40	C = 96
Are you satisfied with your decision? (Y/N)				

Name of Stock	Game Number 5 Trial Number 3		Number of Units	Value of Investment
	Value of Stock	Change		
A	101.60	1.6	60	6096.00
B	82.52	9.4	40	3500.80
C	44.40	-11.2	96	4262.40
Bonds	102.00	2.0	0	0.00
Total Capital:	13,859.20 - 9,180 = 4,679.20			
Information? (Y/N)				

4. The changes in the price of stock C will be sampled randomly from a normal distribution that will remain fixed throughout the experiment. The distribution has a mean of 7 percent and standard deviation of 12 percent (approximately 95 percent of the changes will fall between -17 and 31 percent).

5. (For group 1) The three stocks are mutually independent. The rates of return for one stock have a correlation of zero with the rates of any other stock. (For group 2) The rates of return for stock A are independent of the rates for stock B or the rates for stock C. In other words, the change in the price of stock A on a given trial is not related to the change in the price of stock B or stock C. However, there is a positive dependence between the rates of return of stocks B and C. You can verify it by observing the previous prices of these two stocks. The correlation between the rates of return of stocks B and C is +0.80. (For group 3) The same as for group 2, with "positive" replaced by "negative" and "+0.80" by "-0.80."

6. Your purpose in the experiment is to maximize your payoff in each problem."

## REFERENCES

- Blume, Marshall and Friend, Irwin, "The Asset Structure of Individual Portfolios and Some Implications for Utility Functions," *Journal of Finance*, May 1975, 30, 585-603.
- , Crockett, Jean and Friend, Irwin, Stockownership in the United States: Characteristics and Trends, *Survey of Current Business*, November 1974, 54, 16-40.
- Kahneman, Daniel, Slovic, Paul and Tversky, Amos, eds., *Judgment Under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press, 1982.
- and Tversky, Amos, "Prospect Theory: An Analysis of Decision Under Risk," *Econometrica*, March 1979, 47, 263-91.
- Kroll, Yoram, Levy, Haim and Rapoport, Amnon, "Experimental Tests of the Mean-Variance Model for Portfolio Selection," *Organizational Behavior and Human Decision Processes*, forthcoming 1988.
- Lintner, John, "Security Prices, Risk, and Maximal Gains from Diversification," *Journal of Finance*, December 1965, 20, 587-615.
- Markowitz, Harry, M., "Portfolio Selection,"



- Journal of Finance*, June 1952, 7, 77-91.
- \_\_\_\_\_, *Portfolio Selection: Efficient Diversification of Investments*, New York: Wiley & Sons, 1959.
- \_\_\_\_\_, "Investment for the Long Run: New Evidence for an Old Rule," *Journal of Finance*, December 1976, 31, 1273-86.
- Mossin, Jan, "Equilibrium in a Capital Asset Market," *Econometrica*, October 1966, 34, 768-83.
- Rapoport, Amnon, "Effects of Wealth on Portfolios under Various Investment Conditions," *Acta Psychologica*, February 1984, 55, 31-51.
- Roll, Richard, "A Critique of the Asset Pricing Theory's Tests; Part I: On the Past and Potential Testability of the Theory," *Journal of Financial Economics*, March 1977, 4, 129-76.
- Ross, Stephen A., "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, December 1976, 13, 341-60.
- Sharpe, William F., "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *Journal of Finance*, September 1964, 19, 425-42.
- Tobin, James, "Liquidity Preference as Behavior towards Risk," *Review of Economic Studies*, February 1958, 25, 65-87.
- Zwick, Rami and Rapoport, Amnon, "Relative Gain Maximization in Sequential Three-Person Characteristic Function Games," *Journal of Mathematical Psychology*, September 1985, 29, 333-59.

# Explosive Rational Bubbles in Stock Prices?

By BEHZAD T. DIBA AND HERSCHEL I. GROSSMAN\*

A number of recent studies address the problem of assessing the contributions of market fundamentals and rational bubbles to stock-price fluctuations—see, for example, Olivier Blanchard and Mark Watson, 1982; Robert Flood, Robert Hodrick, and Paul Kaplan, 1986; and Kenneth West, 1986, 1987. A rational bubble reflects a self-confirming belief that an asset's price depends on a variable (or a combination of variables) that is intrinsically irrelevant—that is, not part of market fundamentals—or on truly relevant variables in a way that involves parameters that are not part of market fundamentals. A basic difficulty involved in testing for the existence of rational bubbles, pointed out by Flood and Peter Garber, 1980, and emphasized by James Hamilton and Charles Whiteman, 1985, is that the contribution of hypothetical rational bubbles to asset prices would not be directly distinguishable from the contribution to market fundamentals of variables that the researcher cannot observe. For example, as Hamilton, 1986, shows, a researcher who is unable to observe or to infer changes in the expectations of market participants, especially if they involve the probable future occurrence of relevant events that are infrequent and discrete, might falsely conclude that rational bubbles exist. In the present context, the probabilities that investors attach to possibilities for future tax treatment of dividend income could act like such an unobservable variable.

Diba and Grossman, 1984, and Hamilton and Whiteman, 1985, propose an empirical strategy based on stationarity tests for obtaining evidence against the existence of explosive rational bubbles without precluding the possible effect of unobservable variables on market fundamentals. The present paper implements such tests for explosive rational bubbles in stock prices using a model that assumes a constant discount rate, but that allows unobservable variables to affect market fundamentals and also allows different valuations of expected capital gains and expected dividends. If the first differences of the unobservable variables and the first differences of dividends are stationary (in the mean) and if rational bubbles do not exist, then the model implies that first differences of stock prices are stationary. The model also implies, using an argument adapted from John Campbell and Robert Shiller, 1987, that, if the levels of the unobservable variables and the first differences of dividends are stationary, and if rational bubbles do not exist, then stock prices and dividends are cointegrated of order (1,1).

These theoretical results do not imply that the finding that first differences of stock prices are nonstationary, or that stock prices and dividends are not cointegrated, would establish the existence of rational bubbles. A finding that stock prices and dividends are not cointegrated could result from the nonstationarity of the unobservable variables in market fundamentals, and a finding that stock-price changes are nonstationary could result from the nonstationarity of changes in these unobservable variables. Such findings also could arise from the inappropriateness of the implicit assumption that dividends are generated by an ARIMA process.

The converse inference, however, is possible. That is, evidence that first differences of stock prices have a stationary mean and/or evidence that stock prices are cointegrated with dividends would be evidence against

\*Research Department, Federal Reserve Bank of Philadelphia, Philadelphia, PA 19106, and Department of Economics, Brown University, Providence, RI 02912, respectively. The views expressed are solely those of the authors and do not necessarily represent the views of the Federal Reserve Bank of Philadelphia or of the Federal Reserve System. We thank John Campbell, Robert Shiller, and anonymous referees for helpful comments on earlier versions of this paper.

the existence of rational bubbles. Except by extremely unlikely coincidence, misspecification of market fundamentals could not offset the contribution of a nonstationary rational bubble to stock prices. In addition to analyzing the stationarity properties of the observed time-series of real stock prices and dividends, this paper also examines the stationarity properties of simulated time-series of hypothetical rational bubbles to determine whether the stationarity tests can detect the relevant nonstationarity when it is present.

Because it looks for evidence against the existence of rational bubbles, the analysis in the present paper, in contrast to the strategy for finding rational bubbles suggested by West, 1986, 1987, does not require the specification of a true difference equation relating stock prices only to other observable variables. West observes that, if we could find such a true difference equation, and if the data rejected the implied market-fundamentals solution for stock prices, then we could conclude that rational bubbles exist. The problem with this approach is that diagnostic tests—as reported, for example, by Flood et al., 1986—reject the difference equations linking stock prices to dividends implied by a constant discount rate as well as by extended models that relate the discount rate to the intertemporal marginal rate of substitution or that incorporate different valuations for capital gains and dividends. These results underscore the need for an empirical strategy that does not preclude the possibility that market fundamentals for stock prices depend on unobservable variables in addition to dividends.

### I. The Model

The theoretical model consists of a single equation that relates the current stock price to the present value of next period's expected stock price and dividend payments and to an unobservable variable—that is,

$$(1) \quad P_t = (1+r)^{-1} E_t(P_{t+1} + \alpha d_{t+1} + u_{t+1}),$$

where

$P_t$  is the stock price at date  $t$  relative to

a general index of prices of goods and services;

$r$  is a constant real interest rate that is appropriate for discounting expected capital gains;

$E_t$  is the conditional expectations operator;

$\alpha$  is a positive constant that values expected dividends relative to expected capital gains;

$d_{t+1}$  is the real before-tax dividend paid to the owner of the stock between dates  $t$  and  $t+1$ ; and

$u_{t+1}$  is a variable that market participants either observe or construct, but that the researcher does not observe.

(As suggested above, this unobservable variable could involve the probabilities that investors attach to possibilities for future tax treatment of dividend income.) If  $\alpha$  were equal to unity and  $u_{t+1}$  were equal to zero for all  $t$ , equation (1) would state that the expected real rate of return from holding equity, including expected dividends and expected capital gains, equals the constant  $r$ . The information set of market participants at date  $t$  on which  $E_t$  is based contains at least the current and past realizations of  $P_t$ ,  $d_t$ , and  $u_t$ .

Equation (1) is a first-order expectational difference equation. Because the eigenvalue,  $1+r$ , is greater than unity, the forward-looking solution for the stock price involves a convergent sum, as long as  $E_t(\alpha d_{t+j} + u_{t+j})$  does not grow with  $j$  at a geometric rate equal to or greater than  $1+r$ . This forward-looking solution, denoted by  $F_t$  and referred to as the market-fundamentals component of the stock price, is

$$(2) \quad F_t = \sum_{j=1}^{\infty} (1+r)^{-j} E_t(\alpha d_{t+j} + u_{t+j}).$$

With  $\alpha$  equal to unity and  $u_t$  equal to zero for all  $t$ , equation (2) would say that the market-fundamentals component of the stock price equals the present value of expected real dividends discounted at the constant rate  $r$ .

The general solution to equation (1) is the sum of the market-fundamentals component,

$F_t$ , and the rational bubbles component,  $B_t$ —that is,

$$(3) \quad P_t = B_t + F_t,$$

where  $B_t$  is the solution to the homogeneous expectational difference equation

$$(4) \quad E_t B_{t+1} - (1+r)B_t = 0.$$

A nonzero value of  $B_t$  would reflect the existence of a rational bubble—that is, a self-confirming belief that the stock price does not conform to the market-fundamentals component,  $F_t$ .

Solutions to equation (4) satisfy the stochastic difference equation

$$(5) \quad B_{t+1} - (1+r)B_t = z_{t+1},$$

where  $z_{t+1}$  is a random variable (or combination of variables) generated by a stochastic process that satisfies

$$(6) \quad E_{t-j} z_{t+1} = 0 \quad \text{for all } j \geq 0.$$

The key to the relevance of equation (5) for the general solution of  $P_t$  is that equation (4) relates  $B_t$  to  $E_t B_{t+1}$ , rather than to  $B_{t+1}$  itself as would be the case in a perfect-foresight model.

The random variable  $z_{t+1}$  is an innovation, comprising new information available at date  $t+1$ . This information can be intrinsically irrelevant—that is, unrelated to  $F_{t+1}$ —or it can be related to truly relevant variables, like  $d_{t+1}$ , through parameters that are not present in  $F_{t+1}$ . The only critical property of  $z_{t+1}$ , given by equation (5), is that its expected future values are always zero.

Diba and Grossman, 1988, review and extend theoretical arguments for ruling out rational stock-price bubbles on the basis of the nonnegativity of stock prices and the optimizing decisions of asset holders. George Evans, 1985, develops another theoretical argument for ruling out rational bubbles by requiring that equilibrium rational expectations solutions to the model should be stable in the sense that, given a small disequilibrium deviation from rational expectations,

the system should return to rational expectations equilibrium under a natural revision rule. The empirical analysis developed in the present paper complements these theoretical analyses.

## II. Stationarity of Stock Prices and Dividends

Consider the market-fundamentals component of the stock price given by equation (2). Assume that the process generating  $d_t$  is nonstationary in levels, but that first differences of  $d_t$  and  $u_t$  are stationary. Then, if rational bubbles do not exist, stock prices are nonstationary in levels but stationary in first differences.

If, however, stock prices contain a rational bubble, then for simple specifications of the process generating  $z_t$ , differencing stock prices a finite number of times would not yield a stationary process. Specifically, from equation (5), first differences of a rational bubble would have the generating process

$$(7) \quad [1 - (1+r)L](1-L)B_t = (1-L)z_t,$$

where  $L$  denotes the lag operator. For example, if  $z_t$  is white noise, then an ARMA process that is neither stationary nor invertible generates  $(1-L)B_t$ . (The only exceptions to nonstationarity discussed in the literature involve rational bubbles that almost surely would burst at a finite future date, as in the specifications of Blanchard, 1979, and Blanchard and Watson, 1982. Such a rational bubble would have innovations with infinite variance, but, as Danny Quah, 1985, demonstrates, it also would have a stationary unconditional mean of zero.)

Allan Kleidon, 1986, analyzes the stationarity properties of stock prices, dividends, and their first differences for Data Set 1 in Robert Shiller, 1981. The work of Blanchard and Watson, 1982; Flood et al., 1986; and West, 1986, 1987, also uses this data set. The price series is Standard & Poor's Composite Stock Price Index for January of each year from 1871 to 1986 divided by the wholesale price index for that month. The dividend series is total dividends accruing to this portfolio of stocks for the calendar year divided by the average whole-

TABLE 1—SAMPLE AUTOCORRELATIONS OF REAL STOCK PRICES, DIVIDENDS, AND THEIR FIRST DIFFERENCES

Number of Lags Series	1	2	3	4	5	6	7	8	9	10
$P_t$	0.94	0.87	0.84	0.79	0.74	0.68	0.63	0.57	0.51	0.45
$d_t$	0.95	0.88	0.82	0.78	0.74	0.70	0.65	0.62	0.59	0.56
$\Delta P_t$	0.06	-0.24	0.12	0.17	-0.00	-0.12	0.15	0.00	-0.07	-0.05
$\Delta d_t$	0.23	-0.16	-0.07	-0.03	-0.01	-0.01	-0.17	-0.13	0.06	0.14

Note: The price ( $P_t$ ) and dividend ( $d_t$ ) series contain 116 observations. Their first differences ( $\Delta P_t$  and  $\Delta d_t$ ) contain 115 observations.

TABLE 2—DICKEY-FULLER TEST RESULTS: NO LAGS

$x_t$	$P_t$	$d_t$	$\Delta P_t$	$\Delta d_t$
$\hat{\mu}$	0.0058 (0.0166)	0.0007 (0.0005)	0.0002 (0.0168)	0.0001 (0.0004)
$\hat{\gamma}$	0.0006 (0.0003)	0.00003 (0.00001)	0.0001 (0.0003)	0.000001 (0.000006)
$\hat{\rho}$	0.90 (0.04)	0.87 (0.05)	0.06 (0.10)	0.23 (0.10)
Standard Error of Estimate	0.071	0.002	0.072	0.002
$\Phi_3$	2.55	3.43	42.38	30.89

Note: Regressions are of the form  $x_t = \mu + \gamma t + \rho x_{t-1} + \text{residual}$ . "Standard errors" are in parentheses below coefficients. Sample size is 100 in all cases. The statistic  $\Phi_3$ , calculated like the  $F$ -statistic, tests the null hypothesis  $(\gamma, \rho) = (0, 1)$  against the alternative  $(\gamma, \rho) \neq (0, 1)$ . The rejection region is the set of values of  $\Phi_3$  above 5.47 (6.49) for a test of size 0.10 (0.05).

sale price index for the year. Tables 1, 2, and 3 report results similar to Kleidon's results.

Table 1 presents sample autocorrelations of these real stock prices and dividends, and their first differences, for one through ten lags. The autocorrelations of the undifferenced price and dividend series both drop off slowly as lag length increases, suggesting nonstationary means. Their patterns correspond closely to what would be expected for integrated moving average processes according to a formula presented by Dean Wichern, 1973. In contrast, autocorrelations of the differenced series, both for prices and dividends, are consistent with the assumption that these series have stationary means. Thus the autocorrelation patterns suggest that the nonstationarity of real stock prices is attributable to their market-fundamentals component and that explosive rational bubbles do not exist in stock prices.

Tables 2 and 3 report Dickey-Fuller, 1981, tests for unit roots in the autoregressive representations of real stock prices, dividends, and their first differences. For each time-series, the estimated OLS regression is

$$(8) \quad x_t = \mu + \gamma t + \rho x_{t-1} + \sum_{i=1}^k \beta_i \Delta x_{t-i} + \text{residual},$$

where  $\Delta$  is the difference operator. The tables report the statistic  $\Phi_3$  of David Dickey and Wayne Fuller, 1981, which is calculated as one would calculate the  $F$ -statistic for  $(\gamma, \rho) = (0, 1)$ . The regressions in Table 2 set  $k$  equal to zero to test the null hypothesis that  $x_t$  follows a random walk with drift against the general alternative  $(\gamma, \rho) \neq (0, 1)$ . The regressions in Table 3 set  $k$  equal to four and, thereby, allow  $\Delta x_t$  to follow an AR(4) process. Each regression discards the

TABLE 3—DICKEY-FULLER TEST RESULTS:FOUR LAGS

$x_t$ :	$P_t$	$d_t$	$\Delta P_t$	$\Delta d_t$
$\mu$	0.0046 (0.0159)	0.0008 (0.0005)	-0.0009 (0.0164)	0.0001 (0.0004)
$\gamma$	0.0007 (0.0003)	0.00003 (0.00001)	0.0001 (0.0002)	-0.000001 (0.000006)
$\rho$	0.88 (0.05)	0.83 (0.06)	0.17 (0.24)	0.03 (0.21)
$\beta_1$	0.16 (0.10)	0.35 (0.10)	-0.07 (0.22)	0.25 (0.19)
$\beta_2$	-0.15 (0.10)	-0.14 (0.11)	-0.31 (0.19)	0.01 (0.16)
$\beta_3$	0.21 (0.10)	0.09 (0.10)	-0.14 (0.14)	0.05 (0.13)
$\beta_4$	0.17 (0.10)	0.04 (0.10)	-0.04 (0.11)	-0.01 (0.10)
Standard Error of Estimate	0.068	0.002	0.070	0.002
$\Phi_3$	3.12	4.42	6.41	10.45

Note: Regressions are of the form  $x_t = \mu + \gamma t + \rho x_{t-1} + \sum_{i=1}^4 \beta_i \Delta x_{t-i}$  + residual. "Standard errors" are in parentheses below coefficients. Sample size is 100 in all cases. The statistic  $\Phi_3$ , calculated like the  $F$ -statistic, tests the null hypothesis  $(\gamma, \rho) = (0, 1)$  against the alternative  $(\gamma, \rho) \neq (0, 1)$ . The rejection region is the set of values of  $\Phi_3$  above 5.47 (6.49) for a test of size 0.10 (0.05).

first few observations to adjust sample size to 100. The rejection region, from Dickey and Fuller's Table VI, is the set of values of  $\Phi_3$  above 5.47 (6.49) for a test of size 0.10 (0.05).

For the undifferenced time-series of real stock prices and dividends, the statistic  $\Phi_3$  does not reject the null hypothesis  $(\gamma, \rho) = (0, 1)$ . For both of the differenced series, the statistic rejects the null hypothesis. The rejections are stronger in Table 2 than in Table 3, probably because the regressions of Table 3 include the regressors  $\Delta x_{t-i}$ , which in most cases do not have significant coefficients and, consequently, reduce the power of the unit root test.

The results reported in Tables 2 and 3 support the impression, based on the sample autocorrelations in Table 1, that both real stock prices and dividends are nonstationary in levels but stationary in first differences. For sample sizes of 100, unit root tests have low power against alternatives slightly less than unity—see, for example, G. B. A. Evans and N. E. Savin, 1984. Accordingly, we can-

not have much faith in the result that the undifferenced series are nonstationary and not borderline stationary. The critical finding for our purposes, however, is that, contrary to what the existence of explosive rational bubbles would imply, the data strongly reject the null hypothesis of a nonstationary mean for first differences of real stock prices. In fact, point estimates of  $\rho$  for the  $\Delta P_t$  regressions of Tables 2 and 3 do not differ significantly from zero.

### III. Cointegration of Stock Prices and Dividends

Rearranging terms in equation (2) and substituting the resulting expression for  $F_t$  into equation (3) yields

$$\begin{aligned}
 (9) \quad P_t - \alpha r^{-1} d_t &= B_t + \alpha r^{-1} \left[ \sum_{j=1}^{\infty} (1+r)^{1-j} E_t \Delta d_{t+j} \right] \\
 &\quad + \sum_{j=1}^{\infty} (1+r)^{-j} E_t u_{t+j}.
 \end{aligned}$$

If the unobservable variable in market fundamentals,  $u_t$ , is stationary in levels, if dividends are first-difference stationary, and if rational bubbles do not exist, then the sum given by the right-hand side of equation (9) is stationary. Thus, although  $P_t$  and  $d_t$  are nonstationary, their linear combination  $P_t - \alpha r^{-1} d_t$ , given by the left-hand side of equation (9), is stationary.

Clive Granger and Robert Engle, 1987, define the components of a vector  $y_t$  of time-series to be cointegrated of order  $(d, b)$  if all components of  $y_t$  are integrated of order  $d$ —that is, have a stationary, invertible, nondeterministic ARMA representation after differencing  $d$  times—and if there exists a vector  $\delta$ , other than the null vector, such that  $\delta'y_t$  is integrated of order  $d - b$  for some  $b > 0$ . They call  $\delta$  the cointegrating vector. Using their terminology, equation (9) says that if the processes generating  $\Delta d_t$  and  $u_t$  are stationary and if  $B_t$  equals zero, then  $P_t$  and  $d_t$  are cointegrated of order  $(1, 1)$  with cointegrating vector  $(1, -\alpha r^{-1})$ .

Drawing on the work of James Stock, 1987, Granger and Engle develop tests for cointegration that involve obtaining an estimate of the cointegrating vector from a cointegrating regression and then applying tests for stationarity to the residuals from this regression. For a test of stationarity of the left-hand side of equation (9), the cointegrating regression would be the OLS regression of  $P_t$  on  $d_t$ .

One test for stationarity of residuals suggested by Granger and Engle would reject the null hypothesis of no cointegration if the Durbin-Watson statistic of the cointegrating regression exceeds the critical values they tabulate. Another test suggested by Granger and Engle involves estimating Dickey-Fuller regressions of the form

$$(10) \quad \Delta v_t = -\rho v_{t-1} + \sum_{i=1}^k \beta_i \Delta v_{t-i} \\ + \text{residual,}$$

on the residuals  $v_t$  of the cointegrating regression. Granger and Engle tabulate the critical values for statistics denoted  $\xi_2$  and

$\xi_3$ , calculated analogously to  $t$ -ratios for  $\rho$  in equation (10), with  $k$  set equal to zero for  $\xi_2$  and to four for  $\xi_3$ .

Estimation of the cointegrating regression of  $P_t$  on  $d_t$  yields a point estimate for  $\alpha r^{-1}$  of 30.50 and a Durbin-Watson statistic of 0.61, which is above the 1 percent critical value of 0.51. John Campbell and Shiller, 1987, also estimate such a cointegrating regression and calculate Granger and Engle's  $\xi_2$  and  $\xi_3$  statistics. They find that the statistic  $\xi_2$  rejects the null hypothesis of no cointegration at the 5 percent level, but the statistic  $\xi_3$  (narrowly) fails to reject even at the 10 percent level.

The results of cointegration tests for  $P_t$  and  $d_t$ , thus, are mixed. The Durbin-Watson statistic rejects the null hypothesis that  $P_t$  and  $d_t$  are not cointegrated at the 1 percent level, the statistic  $\xi_2$  rejects the null at the 5 percent level, but the statistic  $\xi_3$  fails to reject at the 10 percent level. Moreover, the point estimate for  $\alpha r^{-1}$  is somewhat implausible. Specifically, with  $\alpha$  set equal to unity, this point estimate would imply a value for  $r$  of about 0.033, well below its sample mean of about 0.08. If  $\alpha$  is less than unity, then the implied value of  $r$  will be even lower than 0.033. (As Terry Marsh and Robert Merton, 1983, emphasize, if the logarithms of stock prices and dividends follow integrated stochastic processes, then a regression of stock prices on dividends, in levels, yields inefficient and possibly biased estimates. This bias could account for the implausibly low values of the required rate of return implied by the cointegrating regression of  $P_t$  on  $d_t$ .)

#### IV. Stationarity of the Unobservable Variable

Nonstationarity of the unobservable variable in market fundamentals would be a potential source of lack of cointegration of stock prices and dividends. To explore this possibility, note that equation (1) implies

$$(11) \quad P_{t+1} + \alpha d_{t+1} - (1+r)P_t = e_{t+1} - u_{t+1},$$

where  $e_{t+1} = P_{t+1} + \alpha d_{t+1} + u_{t+1}$

$$- E_t(P_{t+1} + \alpha d_{t+1} + u_{t+1}).$$

TABLE 4—BHARGAVA TESTS OF THE RANDOM-WALK HYPOTHESIS

Statistic	$R_1$	$R_2$	$N_1$	$N_2$
Null Hypothesis	Random Walk	Random Walk with Drift	Random Walk	Random Walk with Drift
Alternative Hypothesis	Stationary	Stationary	Unstable	Unstable
Rejection Region for test of size 0.05	Above 0.26	Above 0.35	Below 0.006	Below 0.022
$P_t - d_t/0.01$	0.15	0.19	0.05	0.19
$P_t - d_t/0.02$	0.40	0.45	0.12	0.64
$P_t - d_t/0.03$	0.62	0.60	0.31	1.11
$P_t - d_t/0.04$	0.48	0.49	0.44	0.97
$P_t - d_t/0.05$	0.35	0.38	0.35	0.76
$P_t - d_t/0.06$	0.28	0.32	0.26	0.62
$P_t - d_t/0.07$	0.24	0.27	0.21	0.53
$P_t - d_t/0.08$	0.21	0.25	0.18	0.47

Note: The statistics  $R_1$ ,  $R_2$ ,  $N_1$ , and  $N_2$  are von Neumann-type ratios that yield most powerful invariant tests of the random-walk hypothesis against one-sided stationary and explosive alternatives.

Because the assumption of rational expectations implies that  $e_{t+1}$  is not serially correlated, stationarity of the left-hand side of equation (11) is equivalent to stationarity of  $u_{t+1}$ . (Of course, in a finite sample, even if  $u_{t+1}$  is nonstationary, the left-hand side of equation (11) can appear stationary if most of its variability results from movements in the forecast error  $e_{t+1}$ .) Stationarity of the left-hand side of equation (11) implies that the variables  $P_{t+1} + \alpha d_{t+1}$  and  $P_t$  are cointegrated of order (1,1) with cointegrating vector  $[1, -(1+r)]$ .

For the present data, the tests suggested by Granger and Engle find cointegration between  $P_{t+1} + \alpha d_{t+1}$  and  $P_t$  for values of  $\alpha$  between 0.5 and 2, which correspond to varying the valuation for dividends from one-half to twice the valuation for capital gains. With  $\alpha$  set equal to unity, for example, the Durbin-Watson statistic of the cointegrating regression is 1.82 (well above the 1 percent critical value of 0.51), Granger and Engle's  $\xi_2$  statistic has a value of 8.74 (again above the 1 percent critical value of 4.07), and their statistic  $\xi_3$  has a value of 3.32 (which is below the critical value of 3.77 at the 1 percent level but comfortably rejects the null hypothesis of no cointegration at the 5 percent level).

As Campbell and Shiller, 1987, point out, the difference  $P_t - \alpha r^{-1} d_t$  is equivalent to a linear combination of the variables  $\Delta d_{t+1}$ ,  $\Delta P_{t+1}$ , and  $P_{t+1} - \alpha d_{t+1} - (1+r)P_t$ . Accordingly, the conclusion that  $\Delta d_{t+1}$ ,  $\Delta P_{t+1}$ , and  $P_{t+1} + \alpha d_{t+1} - (1+r)P_t$  are all stationary would imply that  $P_t - \alpha r^{-1} d_t$  is stationary, independently of the model of stock prices. Thus, the apparently mixed results of Section III on the hypothesis that stock prices and dividends are not cointegrated are puzzling. (Using the same test, Campbell and Shiller, 1986, find that the logarithm of the ratio of dividends to stock prices and the logarithm of dividends are stationary, but, contrary to what an algebraic identity would imply, they fail to reject the hypothesis that the logarithm of stock prices is nonstationary.)

## V. Bhargava Tests

Given these problems, alternative tests of the hypothesis that  $P_t - \alpha r^{-1} d_t$  is not stationary seem to be in order. To investigate the stationarity properties of  $P_t - \alpha r^{-1} d_t$ , further, this section reports von Neumann-type ratios, suggested by Alok Bhargava, 1986, that yield most powerful invariant tests of random-walk hypotheses against the one-



TABLE 5—AUTOCORRELATIONS OF FIRST DIFFERENCES OF SIMULATED RATIONAL BUBBLE SERIES

Simulation Number	$r_1$	$r_2$	$r_3$	$r_4$	$r_5$	$r_6$	$r_7$	$r_8$	$r_9$	$r_{10}$
1	0.94	0.89	0.84	0.80	0.75	0.71	0.67	0.63	0.59	0.56
2	0.93	0.89	0.83	0.79	0.75	0.71	0.67	0.63	0.60	0.57
3	0.92	0.87	0.80	0.77	0.73	0.70	0.65	0.62	0.57	0.55
4	0.93	0.87	0.82	0.79	0.75	0.72	0.67	0.63	0.59	0.56
5	0.91	0.85	0.78	0.74	0.70	0.66	0.63	0.60	0.56	0.53
6	0.95	0.90	0.85	0.80	0.76	0.71	0.67	0.64	0.60	0.56
7	0.94	0.89	0.84	0.80	0.76	0.72	0.68	0.63	0.59	0.56
8	0.93	0.88	0.84	0.79	0.76	0.72	0.68	0.64	0.60	0.57
9	0.90	0.85	0.82	0.77	0.73	0.70	0.64	0.61	0.58	0.55
10	0.65	0.65	0.62	0.56	0.53	0.45	0.51	0.43	0.42	0.36
11	0.94	0.88	0.83	0.78	0.75	0.71	0.67	0.62	0.58	0.55
12	0.91	0.85	0.82	0.76	0.73	0.68	0.64	0.61	0.56	0.53
13	0.92	0.87	0.82	0.78	0.74	0.70	0.67	0.63	0.60	0.56
14	0.62	0.64	0.55	0.60	0.51	0.46	0.44	0.43	0.39	0.40
15	0.80	0.80	0.73	0.71	0.65	0.65	0.58	0.52	0.53	0.48
16	0.94	0.89	0.84	0.80	0.75	0.71	0.67	0.63	0.59	0.56
17	0.90	0.86	0.81	0.76	0.72	0.68	0.65	0.61	0.59	0.55
18	0.93	0.89	0.84	0.79	0.75	0.70	0.66	0.62	0.59	0.56
19	0.94	0.89	0.84	0.80	0.76	0.72	0.68	0.64	0.60	0.56
20	0.94	0.89	0.85	0.80	0.76	0.72	0.67	0.64	0.60	0.57
21	0.46	0.42	0.41	0.29	0.35	0.31	0.27	0.22	0.24	0.29
22	0.93	0.88	0.83	0.78	0.74	0.70	0.66	0.62	0.58	0.54
23	0.93	0.88	0.83	0.79	0.74	0.70	0.66	0.62	0.59	0.56
24	0.94	0.90	0.85	0.80	0.75	0.71	0.67	0.63	0.60	0.56
25	0.94	0.89	0.84	0.80	0.75	0.71	0.67	0.64	0.61	0.57
26	0.93	0.88	0.83	0.78	0.74	0.70	0.66	0.63	0.59	0.55
27	0.83	0.78	0.75	0.70	0.68	0.67	0.62	0.58	0.53	0.49
28	0.94	0.89	0.84	0.80	0.75	0.71	0.67	0.64	0.60	0.56
29	0.56	0.53	0.48	0.51	0.47	0.40	0.41	0.35	0.35	0.35
30	0.94	0.89	0.84	0.80	0.75	0.71	0.68	0.64	0.60	0.56
31	0.89	0.84	0.80	0.75	0.72	0.69	0.65	0.61	0.55	0.53
32	0.93	0.88	0.83	0.79	0.74	0.70	0.66	0.61	0.58	0.55
33	0.11	0.17	0.21	0.17	0.19	0.12	0.23	0.02	0.18	0.06
34	0.94	0.89	0.85	0.80	0.75	0.71	0.67	0.63	0.59	0.56
35	0.40	0.30	0.30	0.25	0.24	0.30	0.31	0.16	0.27	0.20
36	0.93	0.89	0.84	0.79	0.75	0.70	0.66	0.63	0.59	0.56
37	0.94	0.90	0.85	0.80	0.75	0.71	0.67	0.64	0.60	0.56
38	0.94	0.89	0.84	0.80	0.76	0.72	0.68	0.64	0.60	0.56
39	0.95	0.90	0.85	0.81	0.76	0.72	0.68	0.64	0.60	0.57
40	0.94	0.89	0.84	0.80	0.76	0.71	0.68	0.64	0.60	0.56
41	0.95	0.90	0.85	0.80	0.76	0.72	0.68	0.64	0.60	0.57
42	0.94	0.89	0.84	0.80	0.75	0.71	0.67	0.64	0.60	0.57
43	0.90	0.86	0.82	0.77	0.73	0.70	0.67	0.61	0.57	0.54
44	0.95	0.90	0.85	0.81	0.76	0.72	0.68	0.64	0.60	0.56
45	0.84	0.81	0.74	0.71	0.66	0.61	0.58	0.56	0.51	0.49
46	0.94	0.89	0.85	0.81	0.76	0.72	0.68	0.64	0.61	0.57
47	0.94	0.89	0.84	0.80	0.76	0.71	0.67	0.63	0.60	0.57
48	0.93	0.89	0.84	0.79	0.76	0.72	0.68	0.64	0.60	0.57
49	0.93	0.89	0.84	0.79	0.74	0.70	0.66	0.62	0.59	0.56
50	0.90	0.85	0.81	0.76	0.73	0.69	0.64	0.58	0.56	0.53

Note: Table reports the autocorrelations of first differences of simulated rational bubbles series:  $B_t = 1.05B_{t-1} + z_t$ , where  $z_t$  is normally distributed white noise, and  $B_0$  is set equal to zero. For each simulation,  $r_k$ ,  $k = 1, \dots, 10$ , is the autocorrelation coefficient at lag  $k$ .

sided stationary and explosive alternatives. Tests against one-sided explosive alternatives are relevant because the existence of explosive rational bubbles would imply that  $P_t - \alpha r^{-1}d_t$  has an explosive, rather than a unit, root.

Table 4 reports the Bhargava tests for  $P_t - \alpha r^{-1}d_t$ . The statistic  $R_1$  rejects the null hypothesis of a simple random walk in favor of the stationary alternative for values of  $\alpha^{-1}r$  between 0.02 and 0.06, and the statistic  $R_2$  rejects the null hypothesis of a random walk with drift in favor of the stationary alternative for values of  $\alpha^{-1}r$  between 0.02 and 0.05. The results of tests based on the statistics  $R_1$  and  $R_2$ , concur with the results of two of the Granger and Engle tests reported above and suggest that  $P_t - \alpha r^{-1}d_t$  is stationary. (The values of  $r$  implied by the tests based on  $R_1$  and  $R_2$ , however, still seem somewhat implausibly low.)

The statistics  $N_1$  and  $N_2$  in Table 4 pertain to testing the null hypotheses that  $P_t - \alpha r^{-1}d_t$  follows either a simple random walk or a random walk with drift against the one-sided explosive alternative. For all values of  $\alpha^{-1}r$ , these statistics fail to reject the null hypothesis that  $P_t - \alpha r^{-1}d_t$  has a unit root. In sum, the Bhargava tests strongly suggest that stock prices and dividends are cointegrated, and, thus, are consistent with the finding that the first differences of stock prices and dividends and any unobservable variable in market fundamentals are all stationary.

#### VI. Stationarity Properties of Simulated Rational Bubbles

To verify that our tests would detect explosive rational bubbles if they were present, we applied the same tests to the time-series of simulated rational bubbles with standard normal innovations. The simulations set  $B_0$  equal to zero and  $r$  equal to 0.05.

The statistic  $N_1$  of Bhargava rejected at the 5 percent level the null hypothesis of a simple random walk in favor of the unstable alternative in 95 out of 100 simulations. For the same 100 simulations, the statistic  $N_2$  rejected, at the 5 percent level, the null hy-

pothesis of a random walk with drift in favor of the unstable alternative in 94 cases.

First differences of the simulated rational bubbles series also exhibited strong signs of nonstationarity. Table 5 reports the sample autocorrelations of the differenced time-series for the first 50 simulations. The patterns of autocorrelation coefficients in all but six cases (simulations numbered 10, 14, 21, 29, 33, and 35) strongly suggest nonstationarity. The autocorrelation function starts at a value of 0.8 or higher and drops off very slowly. For simulations numbered 10, 14, 21, 29, and 35, the starting values are lower, but the autocorrelations still drop off slowly. (Wichern's results indicate that the latter criterion is a more reliable sign of nonstationarity.) Only for simulation number 33 does the pattern of autocorrelations resemble those of differenced time-series of stock prices and dividends reported in Table 1 above.

The simulation results reported above, of course, do not mean that stationarity tests would detect a rational bubbles component even if its contribution to stock-price fluctuations is quantitatively small. If, however, the excess volatility in stock prices found by West, 1986, were attributable to rational bubbles, then innovations in these rational bubbles would account for 80 to 95 percent of the variance of stock-price innovations. It is likely then that the stationarity properties of stock prices and dividends would reflect the existence of explosive rational bubbles.

#### VII. Summary

This paper reports empirical tests for the existence of explosive rational bubbles in stock prices. The analysis focuses on a model that defines market fundamentals to be the sum of an unobservable variable and the expected present value of dividends, discounted at a constant rate, and defines a rational bubble to be a self-confirming divergence of stock prices from market fundamentals in response to extraneous variables. The pattern of autocorrelations in the data as well as Dickey-Fuller tests both indicate that stock prices and dividends are nonsta-

tionary before differencing, but are stationary in first differences. In contrast, first differences of simulated time-series of rational bubbles exhibit strong signs of non-stationarity.

If the nonstationarity of dividends accounts for the nonstationarity of stock prices, then stock prices and dividends are cointegrated. Although application of the cointegration tests suggested by Granger and Engle produced somewhat mixed results, these mixed results probably reflect low power of the tests rather than either the existence of rational bubbles or the presence of a nonstationary unobservable variable in market fundamentals. Most importantly, alternative tests suggested by Bhargava indicate that the relevant linear combination of stock prices and dividends is neither explosive nor has a unit root. In contrast, time-series of simulated rational bubbles failed the Bhargava tests. In sum, the analysis supports the conclusion that stock prices do not contain explosive rational bubbles.

## REFERENCES

- Bhargava, Alok, "On the Theory of Testing for Unit Roots in Observed Time Series," *Review of Economic Studies*, July 1986, 53, 369-84.
- Blanchard, Olivier, "Speculative Bubbles, Crashes, and Rational Expectations," *Economic Letters*, 1979, 3, 387-89.
- and Watson, Mark, "Bubbles, Rational Expectations, and Financial Markets," in *Crises in the Economic and Financial Structure*, P. Wachtel, ed., Lexington: Lexington Books, 1982.
- Campbell, John and Shiller, Robert, "Cointegration and Tests of Present Value Models," *Journal of Political Economy*, October 1987, 95, 1062-88.
- , "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," unpublished paper, October 1986.
- Diba, Behzad and Grossman, Herschel, "Rational Bubbles in the Price of Gold," NBER Working Paper No. 1300, March 1984.
- , "The Theory of Rational Bubbles in Stock Prices," NBER Working Paper No. 1990, revised March 1988.
- Dickey, David and Fuller, Wayne, "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root," *Econometrica*, July 1981, 49, 1057-72.
- Evans, George, "Expectational Stability and the Multiple Equilibria Problem in Linear Rational Expectations Models," *Quarterly Journal of Economics*, November 1985, 100, 1218-33.
- Evans, G. B. A. and Savin, N. E., "Testing for Unit Roots: 2," *Econometrica*, September 1984, 52, 1241-69.
- Flood, Robert and Garber, Peter, "Market Fundamentals Versus Price Level Bubbles: The First Tests," *Journal of Political Economy*, August 1980, 88, 745-70.
- , Hodrick, Robert and Kaplan, Paul, "An Evaluation of Recent Evidence on Stock Market Bubbles," unpublished paper, March 1986.
- Fuller, Wayne, *Introduction to Statistical Time Series*, New York: Wiley & Sons, 1976.
- Granger, Clive and Engle, Robert, "Dynamic Model Specification with Equilibrium Constraints: Cointegration and Error-Correction," *Econometrica*, March 1987, 55, 251-76.
- Hamilton, James, "On Testing for Self-Fulfilling Speculative Price Bubbles," *International Economic Review*, October 1986, 27, 545-52.
- and Whiteman, Charles, "The Observable Implications of Self-Fulfilling Expectations," *Journal of Monetary Economics*, November 1985, 16, 353-73.
- Kleidon, Allan, "Variance Bounds Tests and Stock Price Valuation Models," *Journal of Political Economy*, October 1986, 94, 953-1001.
- Marsh, Terry and Merton, Robert, "Aggregate Dividend Behavior and Its Implications for Tests of Stock Market Rationality," Sloan School of Management Working Paper No. 1475-83, September 1983.
- Quah, Danny, "Estimation of a Nonfundamentals Model for Stock Price and Dividend Dynamics," unpublished paper, September 1985.

**Shiller, Robert**, "Do Stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends?," *American Economic Review*, June 1981, 71, 421-36.

**Stock, James**, "Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors," *Econometrica*, September 1987, 55, 1035-56.

**West, Kenneth**, "Dividend Innovations and

Stock Price Variability," NBER Working Paper No. 1833, February 1986.

\_\_\_\_\_, "A Specification Test for Speculative Bubbles," *Quarterly Journal of Economics*, August 1987, 102, 553-80.

**Wichern, Dean**, "The Behavior of the Sample Autocorrelation Function for an Integrated Moving Average Process," *Biometrika*, August 1973, 60, 235-39.

# About Two Marks: Refugees and the Exchange Rate Before the Berlin Wall

By IRWIN L. COLLIER, JR., AND DAVID H. PAPELL\*

Flexible exchange rates among industrial capitalist economies have been a key element of the current international economic system since 1973. In the search for historical lessons to guide present policy, economists have studied the record of flexible exchange rates in the 1920s and the floating exchange rate between Canada and the United States in the 1950s.<sup>1</sup> Compared to exchange rates between capitalist economies, exchange rates between capitalist and socialist economies have received much less attention. These rates are typically set by administrative fiat in socialist economies and usually coexist with an active domestic black market for foreign currencies.<sup>2</sup>

In this paper we examine the history of an unusual floating exchange rate during a unique period of economic history that, to our knowledge, has never before been the subject of econometric investigation. The trading of the East German mark (EM) for the deutsche mark (DM) of the West is completely legal under West German law and the organized exchange of EM and DM has taken place at and among the exchange parlors (*Wechselstuben*) of West Berlin since August 1948.<sup>3</sup> While the private exchange of

these two currencies in West Berlin presupposes some violation of East German currency laws—at the very least currency must be taken across a border illegally—this market is distinct from the black market for DM in East Germany. What makes the history of this particular East/West mark exchange rate interesting is the enormous mobility of people and their currencies between the two economic systems of Berlin which existed until the erection of the Wall on August 13, 1961.

Previous research on fluctuations in the EM/DM exchange rate has not gone beyond a simple matching of upturns and downturns with particular historical events. In this paper we provide econometric evidence that retail sales in East Germany and the flow of refugees from East to West affected the EM/DM exchange rate in West Berlin and that these effects were statistically significant. Decomposing the variables into anticipated and unanticipated components, we show that unanticipated increases in the flow of refugees caused unexpected depreciation of the East mark.

The paper begins by providing the reader with brief historical and institutional surveys of the marketplace where East mark meets West mark in Section I. We then present documentary evidence in Section II which we have assembled from contemporary press reports to illustrate both the supply and demand sides of the EM market in pre-Wall Berlin. In Section III, this "anecdotal evidence" serves to motivate the econometric tests which follow in Section IV. In Section V we summarize our empirical findings.

## I. Historical Background

The monetary division of postwar Germany lagged its political division by three years of economic chaos. The monetary re-

\*Department of Economics, University of Houston, Houston, TX 77004. We are grateful to Gerald Dwyer, Erich Klinkmüller, Frederic Mishkin, Mark Rush, and two anonymous referees for their helpful comments and suggestions. An earlier version of the paper was presented at the 1986 Winter Econometric Society Meetings.

<sup>1</sup>For example, see Jacob Frenkel, 1980, and Paul Wonnacott, 1972.

<sup>2</sup>For a comprehensive recent survey of the vast array of exchange rates for the Eastern European economies see Jozef van Brabant, 1985.

<sup>3</sup>The official name of the East German currency unit from July 1948 through July 1964 was *Deutsche Mark der Deutschen Notenbank*. Our choice of EM to denote the other German mark, while idiosyncratic to this paper, is notationally convenient.

form of June 22, 1948 in the Western occupational zones heralded the revival of the West German economy as well as the end of the charade that "During the period of occupation Germany shall be treated as a single economic unit," as had been agreed at the Conference in Potsdam.<sup>4</sup> On the following day Soviet Marshal Sokolovsky issued Order No. 111 introducing a currency reform for the Soviet Occupational Zone and Greater Berlin. The first East mark, nicknamed the "Coupon mark," was simply the old Reichsmark pasted with a special coupon. The birth of the twin German marks was immediately followed by the Soviet blockade of the Western sectors of Berlin which was eventually broken by the Western airlift.

In an apparent attempt to expand its control over both the Western and Eastern sectors of Berlin, the Soviet Military Administration of Germany tried to prohibit the introduction or circulation of the deutsche mark in either the Soviet Zone of Occupation or the territory of *Greater Berlin*.<sup>5</sup> Resisting the Soviet claim to sole control over monetary matters regarding Greater Berlin, the Western occupational authorities found themselves forced to link the Western sectors of Berlin to the monetary reform which had been declared for the Western occupational zones.<sup>6</sup>

The original U.S. position was that a freely circulating uniform currency for all of Berlin was desirable. Given anticipated difficulties

in maintaining a separate currency for the Western sectors of Berlin, a monetary union along the lines of Liechtenstein and Switzerland would have been acceptable, provided the issue of the Soviet Zone mark remained under quadripartite agreement and control. However, since the Soviet Army had not been forthcoming with information on the quantity of Allied military marks previously printed in the Soviet Zone (Smith, 1974, p. 303), Clay and his financial advisers planned for the contingency that the Soviets would be unwilling to participate in quadripartite control of money in Berlin and the Soviet Zone.

The First Decree for Monetary Reorganization which went into effect for the Western sectors of Berlin on June 25, 1948, contained an important paragraph, §4a, in which both the DM and the EM ("that currency which is the legal tender in the Soviet sector of Berlin") were declared to have equal value and to be legal tender for purchases in the Western sectors for a set of "basic" goods and services listed in the Decree.<sup>7</sup> For all goods and services not listed in the decree, the seller of goods or services could choose the means of payment. After it became clear which currency was preferred by workers, the amendment to the original monetary reorganization decree of July 4, 1948, declared that no employee was permitted to demand more than 25 percent of wage or salary be paid in Western currency. The reasons for the Western support of a 1:1 parity for the payment of basic goods and public services were political. There was a desire to protect those workers who lived in the Western sectors of Berlin but worked in East Berlin or the Soviet Occupied Zone and paid in East marks. More importantly, the Western allies were still committed to keeping Greater Berlin under a single-municipal administration.

<sup>4</sup>The charade began with the establishment of occupational zones. Both the Soviet and French governments were particularly unwilling to enter into agreements concerning Germany as a whole. A list of specific instances of early Soviet failure to carry out joint agreements is found in Jean Smith, 1974, pp. 243-4.

<sup>5</sup>The Western interpretation of the division of Germany is that Germany had been divided into four zones of occupation and Greater Berlin with the latter in turn subdivided into four occupational sectors. The Soviet interpretation was that Greater Berlin being located in the Soviet Zone represented an economic component of the Soviet Zone.

<sup>6</sup>For an eyewitness account of the four-power negotiations over the currency reform, the reader is referred to the report written from Berlin by General Lucius D. Clay, dated June 25, 1948, in Smith (1974, p. 698).

<sup>7</sup>Specifically mentioned in §4a were bread, potatoes, flour, meat, lard, sugar, malt-coffee, salt, land rent, public transportation, postal services including telephone and telegraph, electricity and gas bills, taxes and other municipal fees.

## II. The West Berlin Exchange Parlors<sup>8</sup>

With the EM legal tender in West Berlin for 75 percent of an employer's wage bill and for household purchases of certain basic goods and services, there was a pressing need for an orderly process for exchanging the two currencies. The banking system was at this time unable to provide this service. For several months following the currency reforms in East and West Berlin the banking and financial institutions of Berlin continued to work undivided. Most of the commercial credit for the Berlin economy immediately after the currency reforms was granted by the Berlin Stadtkontor, which had its main offices in the Soviet sector of the city. Indeed the branch offices of the Stadtkontor located in the Western sectors continued to grant credits in the Eastern currency. At the end of July 1948 the Western allies permitted a special postal check office to be opened in West Berlin. Separate accounts in DM and EM were administered but this office was not recognized by the Soviet Occupational authorities and postal cash transfers were broken off between the West Berlin and Soviet-occupied Germany.

On July 27, 1948, acting upon recommendations from members of the Berlin banking community and the German Currency Commission, the commanders of the three Western sectors of Berlin directed the mayors of the administrative districts of the Western sectors of Berlin to allow their respective departments for trade, handicrafts, and business to permit the establishment of exchange parlors. The procedure for obtaining a license to run an exchange parlor was to follow established procedures for setting up any other new business.

The first exchange parlors were licensed in the British sector and opened for business on August 2, 1948, buying and selling 1 DM for 2.2 EM. In February 1949 there were already

close to 40 exchange parlors open in West Berlin and by the end of the year there were 58.<sup>9</sup> Most of the business of the exchange parlors in this initial period was to satisfy the demands of West Berlin enterprises for EM to be used for paying wages and salaries. The walk-in business of individuals changing small sums of money played only a minor rôle at that time.

Both the collapse of last hopes for a unified city government in November 1948 and the inexorable workings of Gresham's law led to a suspension of §4a on March 20, 1949. Henceforward the deutsche mark was the sole legal tender in the Western sectors of Berlin. There was no longer a demand for EM to finance transactions within West Berlin.

The Currency Commission which had been established by the Allied Kommandatur with the currency reform in June 1948 was transformed into the Berlin Central Bank on March 20, 1949. The permission to establish private banks in West Berlin was granted by the Allied Kommandatur in early July 1949. The West Berlin exchange parlors were only regarded as a transitional solution until banks in West Berlin could be chartered. The exchange parlors were prohibited by law from engaging in other banking operations. However, by the time private banks were chartered in West Berlin, the exchange parlors had established themselves as an integral part of the West Berlin economy.<sup>10</sup>

## III. Supplies and Demands: Anecdotal Evidence<sup>11</sup>

In this section we present illustrative cases from contemporary East and West German

<sup>8</sup> The discussion in this section draws heavily on the very useful unpublished Diplomarbeit of Claus Knetschke (1956), "Der Handel D-Mark-West—D-Mark-Ost." A carbon copy of Knetschke's Diplomarbeit is available in the library of the Gesamtdeutsches Institut in Bonn.

<sup>9</sup> Roswitha Urban, 1949, and *Berlin in Zahlen* (1950, p. 165). According to a recent article in the Hamburg weekly newspaper, *Die Zeit* (North American edition), February 6, 1987, "Der Kurs ging in den Keller" about a half dozen small private exchange parlors are still in business in West Berlin.

<sup>10</sup> The buying and selling of East marks is not subject to West German foreign exchange laws because the Soviet Occupied Zone and later GDR was and still is not regarded as a foreign country.

<sup>11</sup> The newspaper reports we cite in the paper were collected at the extraordinary newspaper clipping archive of the Gesamtdeutsches Institut in West Berlin. We are grateful to Günther Buch for providing us access to the archive.

sources. The supply of EM was dominated by East Germans wanting to shop in West Berlin. Refugees escaping East Germany also brought considerable sums of EM along with them to exchange on their respective ways through West Berlin. In addition there were substantial capital flows in response to political events. Working the supply and demand sides of the market for EM were over 50,000 border-crossing commuters (*Grenzgänger*)—those who lived on one side of the West Berlin border and worked on the other.<sup>12</sup> There were two distinct components of the demand for EM in West Berlin. The first was the demand for EM by West Berlin consumers for whom many East German goods and services at the market exchange rate were bargains. The second component was the demand for monetary transfers from those living in the West to their relatives and friends in the East. We conclude this section of the paper with a brief discussion of the problem of the mismatch of EM denominations bought and sold at the exchange parlors.

*East German Consumers and West Berliner Goods and Services.* The most important source of supply of EM was the average East German consumer wanting to purchase goods and services in West Berlin. With eight rapid transit lines (S-Bahn), four subway lines (U-Bahn), and 193 roads crossing into West Berlin, there was really little the East German authorities could do to control the flows of the EM to and from West Berlin. The act of taking EM to West Berlin for shopping was a violation of GDR law and the penalties for being caught could be severe. For exchanging money to buy 381 EM worth of clothing (a knit jacket, one pair each of corduroy trousers and jeans, and a sweater), one East German citizen was given a three-month jail sentence.<sup>13</sup> Another case reported in the East German press involved a 52-year-old bookkeeper at the HO-Dresden (a retail trade organization) who was

given six weeks in jail for smuggling 200 EM for her shopping trip to West Berlin. Her purchases included 18 chocolate bars, 2 pounds of coffee, a bottle of perfume, and 4 pairs of nylon stockings.<sup>14</sup> The attraction of shopping in West Berlin was strong enough to overcome even the political sensibilities of some wives of army officers<sup>15</sup> and members of East Germany's Socialist Unity Party (SED).<sup>16</sup>

Even though the EM was not legal tender for any class of transactions in West Berlin after March 1949, EM were often accepted at the market exchange rate by shopkeepers and department stores who remained the large-volume customers of the exchange parlors. Because the exchange parlor rate was fixed and known for each day, store owners in West Berlin could accept EM at little risk, knowing exactly what they would receive.<sup>17</sup>

<sup>14</sup>*Berliner Zeitung* (East Berlin), May 11, 1958. What is particularly interesting about both cases is that every item mentioned can now be purchased legally by East Germans using DM vouchers (purchased with DM) in the special *Intershops* found throughout the GDR. From a newspaper story reporting the details of 14 cases of smugglers caught by GDR customs officers one can obtain a sense of the distribution of the amounts smuggled by individuals. For the 14 reported cases, the mean and standard deviation of the value of the West Berlin goods bought by East Germans were 98 EM and 51 EM, respectively. *Freiheit* (Köthen, GDR, July 15, 1960).

<sup>15</sup>The West Berlin daily *Der Tagesspiegel* (July 2, 1958) quoted present-day GDR leader, Erich Honecker, from an article in the East German army newspaper *Die Volksarmee*. The wives were reported to have received stiff sentences.

<sup>16</sup>The district communist party control committee of Potsdam initiated disciplinary proceedings against 189 members and candidates of the SED for having gone to West Berlin and exchanged money. Expulsion from the party was the likely outcome of the proceedings. *IWE* (West Berlin), January 20, 1955. At an SED county leadership meeting for the city of Rostock it was declared that "The time of treating violations of the Law for the Protection of Intergerman Payments as minor transgressions (*Kavaliersdelikte*) has passed once and for all." *Ibid.*, November 25, 1960.

<sup>17</sup>Accepting EM as a West Berlin store owner did entail minor risks of another kind. We were able to find one instance where the names of the owners of a fruit and vegetable store and a leather goods store in West Berlin had been given to East German customs officers who had caught some of their customers coming back to

<sup>12</sup>Considerably more border-crossing commuters lived in the East and worked in West Berlin than vice versa.

<sup>13</sup>*Nationalzeitung* (GDR), May 10, 1955.



Analogous to the present popularity of western television programs in East Germany, movies in West Berlin were a popular source of entertainment for the residents of the East before the Berlin Wall. It was quite typical for the East Berlin public to pay EM for admission to movies showing in West Berlin. Those who had the misfortune of having western movie stubs in their pockets during a customs control could be charged with violations of currency laws.<sup>18</sup>

The demand for DM was not met exclusively with a supply of EM. An extreme, if somewhat macabre, case was that of the 19-year-old East German man who sold his blood in West Berlin for DM. He was sentenced to 2 1/2 years in jail for his crime which technically involved no violation of East German law because "the socialist lawmakers could not have anticipated such an unconscionable abomination."<sup>19</sup> More typical were the East Germans who smuggled conventional stores of value to West Berlin such as gold jewelry and collectibles.

*Refugees.* While it was relatively easy to pass from East Germany to West Berlin before the Wall, it was still prudent for a refugee family not to draw excessive attention to itself by carrying its entire *Hab und Gut* along in the S-Bahn. For all the historical reasons that have made paper money so convenient, it was also an ideal asset for the average refugee to bring along for the necessary portfolio adjustment at the exchange parlor on the way to the emergency reception center at Marienfelde. In contemporary accounts one finds mention of the perceived influence of the magnitude of the refugee flow on the Berlin exchange rate.<sup>20</sup>

the East. The private cars of the West Berlin shop owners were confiscated during a trip to the East and held pending payment of fines. *Neue Zeit* (GDR), April 6, 1957.

<sup>18</sup> Compare *Märkische Volksstimme*, July 17, 1958. One youth caught by vigilant customs officers was alleged to have confessed to having spent about 100 EM during his trips to the West.

<sup>19</sup> *IWE-Tagesdienst* (West Berlin), October 27, 1961, No. 11, 1961.

<sup>20</sup> For example, "...[The exchange rate] has quickly again shown a tendency to fall because even that little

*Border-Crossing Commuters.* One of the striking peculiarities of pre-Wall Berlin was the large number of individuals who continued to live on one side of the border but to commute to jobs on the other side. By 1961 there were over 50,000 border-crossing commuters. West Berlin government policy was to promote such personal links in the politically divided city. There was some financial incentive to live (i.e., spend EM) in the East and earn DM in the West but not vice versa. A system was established in which East-to-West commuters were in effect forced to subsidize those commuting from West to East. A West Berlin commuter who worked in East Berlin was permitted to exchange a certain fraction of his or her EM wage for DM at the very favorable rate of 1:1 through the Wage Adjustment Fund (*Lohnausgleichskasse*) established by the West Berlin city government. The DM in the Wage Adjustment Fund came from the much larger fraction of the DM wages of East Berlin commuters working in West Berlin, which they were legally required to exchange for EM from the fund at 1:1.<sup>21</sup> According to Knetschke (1956), the Wage Adjustment Fund was supposed to balance EM receipts with EM payments and did not buy or sell EM at the exchange parlors.<sup>22</sup> Border-cross-

cash which is brought along by individual refugees from the East creates a large supply [of EM] for the exchange parlors as the aggregate number of refugees grows." "DM-Ost—Wahrheit und Illusion" [DM-East—Truth and Illusion] by Georg Gnieser in *Der Volkswirt*, Vol. 14, No. 19, May 7, 1960. Also Knetschke, 1956, pp. 76, 81.

<sup>21</sup> The required fraction exchanged by the East Berlin commuters dropped from a high of 90 percent in 1949 to 60 percent in 1961. West Berlin commuters were originally permitted to exchange 60 percent of their EM wages in 1949. Apparently that fraction fell to about one-third in the early 1950s.

<sup>22</sup> From the limited information we have found describing the mechanics of the Wage Adjustment Fund, it appears highly unlikely that there would have been a balance between the inflow of EM paid by West Berliners working in the East and the outflow of EM paid to East Germans working in West Berlin unless a significant portion of the DM wages of East German commuters escaped the Wage Adjustment Fund. The reason for this supposition is that there were relatively too few West Berlin commuters working in the East to have

ing commuters were still free to go to the exchange parlors and convert any of the remaining portion of their wages into EM or DM at the market rate.<sup>23</sup>

*Purchases by West Berliners in East Germany.* One indication of what West Berliners were buying in the East can be found in the restrictions upon what they were allowed to buy using EM. For example, it was reported less than two months before the erection of the Berlin wall that food and drink in restaurants or theater buffets would no longer be sold to Western visitors for EM.<sup>24</sup>

A better indicator are the kinds of items which East German customs officials seized at the border to West Berlin. Typical goods which were reported to have been smuggled to the West include butter, meat, cameras, optical instruments, typewriters, Jena glass, art objects, musical instruments, and porcelain. During the approximately eight weeks between March 10 to May 6 of 1958, East German customs officials were reported to have confiscated 29,000 eggs and two metric tons of meat and sausage being illegally exported.<sup>25</sup>

Some interesting (and perhaps even accurate) statistics can be found in the post-wall East German press.<sup>26</sup> Between 1953 and 1960 East German customs were able to stop 318,370 instances of goods being illegally exported with a total value of 9.1 million EM (average value of 29 EM/instance). In 1960 20,608 instances of illegal exports with

a total value of confiscated goods of .98 million EM (48 EM average) were registered by GDR customs.<sup>27</sup> In the first week after August 13, 1961, meat sales in East Berlin were claimed to have dropped 100 metric tons and butter sales to have dropped 35 tons. During the period August 13 to September 26, 1961, the sale of meat and butter allegedly dropped 900 and 300 metric tons, respectively, compared to the same period of the previous year.<sup>28</sup>

*Support Payments from West to East.* Both as a consequence of population losses from the war and the separation of families resulting from the division of Germany, there existed a significant demand for EM to support family and friends in East Germany.<sup>29</sup> One can find many stories in the East German press attacking families in which the husbands worked and lived in West Berlin and who exchanged DM for EM at the exchange parlors to support their wives living in East Germany.<sup>30</sup> Apparently this was not only a practice for intact families. In one instance an East German divorcee received

---

generated sufficient EM in the Wage Adjustment Fund to have covered the full "coerced demand" for EM of the much more numerous East German commuters working in West Berlin.

<sup>23</sup>*Berliner Zeitung* (GDR), February 18, 1959, and *IWE* (West Berlin), February 17, 1959, both reported a public lecture "The Exchange Rate Viewed Scientifically" by a certain Dr. Wemmer of the Humboldt University in East Berlin. Wemmer no doubt gave the official SED line when he characterized the Berlin border-crossing commuters as "legalized agents and smugglers."

<sup>24</sup>*IWE* (West Berlin), June 20, 1961.

<sup>25</sup>*Berliner Zeitung* (East Berlin), May 11, 1958.

<sup>26</sup>The statistics in this paragraph have been taken from *Junge Welt* (GDR), January 6, 1962, and *Berliner Zeitung* (East Berlin), May 1, 1962, p. 11.

<sup>27</sup>These articles are inconsistent in one important respect. Both claim that only 1 in 10,000 persons crossing the border was controlled by customs officers and of those controlled 4 to 5 percent were found to have been illegally exporting East German goods. These figures appear to contradict other figures given in the article—1.5 million average daily crossings and 318,370 instances of smuggling found by customs officers.

<sup>28</sup>According to the 1962 (East) Berlin Statistical Yearbook (*Statistisches Jahrbuch Berlin 1962*, p. 288), on a per capita basis 67.5 kilograms of meat, sausage, and meat products and 14.1 kilograms of butter were supplied to East Berlin retail establishments during 1961. The population of East Berlin was 1.055 million at the end of 1961. Thus the annual figures for the supply of meat and butter for East Berlin in 1961 would be 71,213 metric tons and 14,876 metric tons, respectively. Thus for an average 6-week period in 1961 East Berlin supplies of meat and butter were about 8217 and 1716 metric tons, respectively. Hence the alleged 900 and 300 metric tons drops in meat and butter sales would have been quite substantial.

<sup>29</sup>"If one examines the motives of this money smuggling, then it is mostly a case of a pension or support payment which is received in West Berlin, exchanged there at the swindle rate, and then—having grown by a factor of four—taken back home to the GDR." *Märkische Volksstimme* (GDR), February 5, 1961.

<sup>30</sup>For example, *Märkische Volksstimme* (GDR), March 7, 1956.

50 DM monthly from her ex-husband, paid into her West Berlin bank account from 1948 to 1956.<sup>31</sup>

According to the East German Law for the Regulation of Intergerman Payments of December 15, 1950, claims by inhabitants of East Germany on West German citizens had to be paid in DM to the *Deutsche Notenbank* (GDR). The Deutsche Notenbank would then convert the DM sum into EM at the official exchange rate of 1:1 for payment to the East German claimant.<sup>32</sup> For some time it was still feasible to make voluntary transfers (for example, those not ordered by a court for alimony or child support) to East German citizens by paying at many banks and financial institutions in West Berlin in DM and having that money converted at the daily West Berlin exchange rate for transfer to an East German bank for payment in EM. This channel was closed in early January 1955.<sup>33</sup> Then the only way to legally transfer EM to someone in East Germany was by purchasing EM at the unfavorable, official East German rate of 1:1. Still most folks preferred going to the West Berlin exchange parlors and accepting the risks of smuggling EM into the GDR than exchanging DM for EM at parity.

*Capital Flows.* There appear to have been substantial capital flows from East Germany to West Berlin which were affected by changes in the expected temperature of the Cold War. According to contemporary accounts, West Berlin bank balances of individuals living in the East rose and fell with the level of the intensity of the Berlin question.<sup>34</sup> One would also expect these capital

flows to be affected by political events such as the Hungarian revolution.

The marketplace for the EM also attracted nonprivate sources of supply. Early on the Soviet Military Administration in Germany (SMAD) was suspected by General Clay of exchanging EM for DM in order "to finance communist activities in western Germany."<sup>35</sup> Later official East German sources must have also been involved in EM supply since it was hardly an uncommon event for suitcases full of newly printed currency to arrive at the exchange parlors. Little more can be said about official EM supply to the West Berlin exchange parlors other than it was certainly one component of the error term in the specifications we have estimated.

*Professional Changers.* An interesting problem which naturally arose in the exchange of EM and DM in West Berlin was the fundamentally different mix of denominations of EM supplied and demanded. Typically, large denomination EM notes were sold to the exchange parlors whereas buyers of EM preferred smaller denominations. The market solution to this problem was the professional changer. According to one 1955 East German newspaper account a 58-year-old man was caught on the border with 50,000 EM in small bills. He was charged with regularly bringing large EM bills from West Berlin which he changed for smaller bills in East Berlin. Over a three-year period he had personally changed an estimated 18 million EM for which he received 25,000 marks.<sup>36</sup>

<sup>31</sup>The newspaper story went on to report that the divorcee's mother had also received money (DM) from her son via a West Berlin friend. *Volksstimme/Magdeburg* (GDR), May 16, 1958.

<sup>32</sup>The official exchange rate quoted by the Deutsche Notenbank was strictly a one-way proposition. The bank was not obligated to convert EM into DM at the 1:1 rate.

<sup>33</sup>*Der Tagesspiegel* (West Berlin), January 14, 1955.

<sup>34</sup>"Up to the beginning of the new stage in the war of nerves concerning Berlin, many of the residents of the zone [authors' note: the GDR and East Berlin is meant] were actively building up West mark accounts

under assumed names or the names of friends and in quite a few instances accumulating considerable sums... In the meantime numerous clients living in the East have closed their West accounts and have changed back into East marks." *Deutsche Zeitung und Wirtschaftszeitung* (FRG), March 11, 1959.

<sup>35</sup>Smith (1974), p. 912.

<sup>36</sup>*Neue Zeit* (GDR), June 17, 1955. Unfortunately it is not clear from the article whether the man received his commission in EM or DM. We presume he kept 25,000 EM which would have been exchanged for DM at the exchange parlors. In the same story it was reported that independent of the previous changer just described, three East Berlin Stadtkontor employees had accepted bribes of 58,200 EM for changing 46.2 million EM. They were sentenced to 3 to 6 years in prison.

#### IV. Econometric Results

The anecdotal evidence presented in the previous section indicates that the East/West mark exchange rate was determined by conventional and unconventional sources of supply and demand in a well organized financial market. We condense the evidence above into two propositions:

**PROPOSITION 1:** *When the flow of refugees from East to West Berlin is high (low), the East mark will depreciate (appreciate).*

**PROPOSITION 2:** *When sales of consumer goods in East Berlin are high (low), the East mark will appreciate (depreciate).*

These propositions are consistent with the predictions of the portfolio balance model of exchange rate determination. Both a high flow of refugees from East to West Berlin and low sales of consumer goods in East Berlin raise the desired proportion of West marks in residents' portfolios, causing the East mark to depreciate. The propositions are also consistent with a partial equilibrium view of the exchange parlor market, where the exchange rate is determined by the supply and demand for the two currencies. A high flow of refugees increases the supply of East marks while low sales (i.e., supplies) of consumer goods in East Germany simultaneously increases the Eastern demand for West marks and lowers the Western demand for East marks. Again high refugee flows and low GDR retail sales are expected to cause a depreciation of the East mark.

Ideally one would test these propositions within one of these structural models of exchange rate determination, but the usual determinants of such models—relative prices, incomes, interest rates, money supplies, and wealth—are either not available or not particularly meaningful for this market. Instead we estimate vector autoregressions for the variables suggested by Propositions 1 and 2 and assess the significance of these variables by Granger (1969) causality tests ( $F$ -tests). This approach enables us to establish the influence of refugees and sales on the exchange rate without specifying a particular

model. We then use these autoregressions to decompose movements of the exchange rate into anticipated and unanticipated components. We find that unanticipated flows of refugees caused unanticipated depreciations of the East mark.

Our sample period runs from February 1954 through July 1961.<sup>37</sup> The EM/DM exchange rate used for the paper is the end-of-month rate as set by the exchange parlors and reported in the West Berlin daily newspaper *Der Tagesspiegel*. The exchange rate ( $E$ ) used in the autoregressions is the first difference of the natural logarithm of the exchange rate. As is usual with exchange rate data, first differencing was necessary in order to achieve stationarity. The total number of refugees registering at emergency reception centers in West Berlin and the Federal Republic of Germany during the month is our refugees' variable. As a measure of the availability of consumption goods in the GDR we use monthly retail sales in East Germany (data for East Berlin were not available for a sufficiently long period).<sup>38</sup> Our refugees ( $R$ ) and retail sales ( $S$ ) variables are natural logarithms of the original levels. Detailed information about our data set is provided in the Data Appendix.<sup>39</sup>

We chose six lags for the autoregressions, after estimating models ranging from 1 to 12 lags, on the basis of both the Akaike Information Criterion and a likelihood ratio test.<sup>40</sup> A constant term is included in all the regres-

<sup>37</sup>The sample period was defined by two "regime" shifts: the uprising of East German workers in June 1953 (plus half a year for lagged terms) and the building of the Berlin Wall in August 1961.

<sup>38</sup>Data on monthly retail sales in East Berlin were only available beginning January 1959. For the period January 1959 through December 1961 the simple correlation coefficient between the levels of retail sales in East Berlin and for the entire GDR was .95.

<sup>39</sup>We were also able to collect and test variables for East German currency in circulation and for production of consumer goods. Neither variable proved to be statistically significant when added to our refugee or sales variables nor did they perform better than the sales variable alone.

<sup>40</sup>These procedures are described in Richard Baillie, Robert Lippens, and Patrick McMahon (1983).

TABLE 1—*F*-TESTS FOR THE INFLUENCE OF REFUGEES AND SALES ON THE EM/DM EXCHANGE RATE

Unconstrained Regression	Constrained Regression	<i>F</i> -Statistic
(1) $E_t = \alpha_0 + \sum_{i=1}^6 \alpha_i E_{t-i} + \sum_{i=1}^6 \beta_i R_{t-i}$	$E_t = \alpha_0 + \sum_{i=1}^6 \alpha_i E_{t-i}$	3.03 <sup>a</sup>
(2) $E_t = \alpha_0 + \sum_{i=1}^6 \alpha_i E_{t-i} + \sum_{i=1}^6 \gamma_i S_{t-i}$	$E_t = \alpha_0 + \sum_{i=1}^6 \alpha_i E_{t-i}$	4.18 <sup>a</sup>
(3) $E_t = \alpha_0 + \sum_{i=1}^6 \alpha_i E_{t-i} + \sum_{i=1}^6 \beta_i R_{t-i}$ $+ \sum_{i=1}^6 \gamma_i S_{t-i}$	$E_t = \alpha_0 + \sum_{i=1}^6 \alpha_i E_{t-i} + \sum_{i=1}^6 \gamma_i S_{t-i}$	4.79 <sup>b</sup>
(4) $E_t = \alpha_0 + \sum_{i=1}^6 \alpha_i E_{t-i} + \sum_{i=1}^6 \beta_i R_{t-i}$ $+ \sum_{i=1}^6 \gamma_i S_{t-i}$	$E_t = \alpha_0 + \sum_{i=1}^6 \alpha_i E_{t-i} + \sum_{i=1}^6 \beta_i R_{t-i}$	6.00 <sup>b</sup>

<sup>a</sup>Marginal significance levels: 2.23 (5%) and 3.07 (1%).<sup>b</sup>Marginal significance levels: 2.24 (5%) and 3.09 (1%).

sions. We have not incorporated monthly dummy variables here because no significant evidence of seasonal patterns was found when they were included.

The central result of the estimation is that we can reject at standard significance levels the null hypothesis that neither refugees nor sales affect (Granger cause) the exchange rate—lines (1) and (2) in Table 1. The *F*-statistic for including refugees in the autoregression is 3.03 and for sales is 4.18, both well above the 5 percent value of 2.23 and, in the case of sales, well above the 1 percent value of 3.07. When the influence of either refugees or sales in addition to the effect of the other is calculated (lines (3) and (4) in Table 1), the results are even stronger. The *F*-statistic for including refugees is 4.79 and for sales is 6.0, both well above the 1 percent level of 3.09.

In Table 2 we show the results of analogous tests which treat refugees and sales as dependent variables. A plausible story for changes in the exchange rate preceding changes in refugee flows would be the relative speed of adjustment of the EM market

compared to migration in response to political changes in the GDR. We have no plausible story for expecting causation to run from the exchange rate in West Berlin to retail sales in the GDR. Fortunately the low *F*-statistics in lines (1) and (3) do not lead to rejection of the null hypothesis that the exchange rate affected neither refugee flows nor sales, indicating that the causality was not bi-directional.

Using our methods, we can test two conventional beliefs regarding refugees and sales. First, one would expect that the losses in labor which necessarily accompanied the westward emigration would have consequences for production and eventually retail sales, that is, increased emigration would be associated with a drop in retail sales. Second, it is generally believed that the emigration decision was in no small part affected by the difference in living standards between the two Germanys—an extended period of lower retail sales (greater shortage) could be expected to increase emigration. The high *F*-statistics in lines (2) and (4) allow us to reject at the 1 percent level both the null

TABLE 2—*F*-TESTS FOR THE INFLUENCE OF THE EM/DM EXCHANGE RATE ON REFUGEES AND SALES

Unconstrained Regression	Constrained Regression	<i>F</i> -Statistic
(1) $R_t = \beta_0 + \sum_{i=1}^6 \beta_i R_{t-i} + \sum_{i=1}^6 \alpha_i E_{t-i}$	$R_t = \beta_0 + \sum_{i=1}^6 \beta_i R_{t-i}$	.76 <sup>a</sup>
(2) $R_t = \beta_0 + \sum_{i=1}^6 \beta_i R_{t-i} + \sum_{i=1}^6 \gamma_i S_{t-i}$	$R_t = \beta_0 + \sum_{i=1}^6 \beta_i R_{t-i}$	4.39 <sup>a</sup>
(3) $S_t = \gamma_0 + \sum_{i=1}^6 \gamma_i S_{t-i} + \sum_{i=1}^6 \alpha_i E_{t-i}$	$S_t = \gamma_0 + \sum_{i=1}^6 \gamma_i S_{t-i}$	1.62 <sup>a</sup>
(4) $S_t = \gamma_0 + \sum_{i=1}^6 \gamma_i S_{t-i} + \sum_{i=1}^6 \beta_i R_{t-i}$	$S_t = \gamma_0 + \sum_{i=1}^6 \gamma_i S_{t-i}$	7.62 <sup>a</sup>

<sup>a</sup>Marginal significance levels: 2.23 (5%) and 3.07 (1%).

hypothesis that refugees did not affect sales and the null hypothesis that sales did not affect refugees, providing support to both conventional beliefs.

While the Granger causality tests indicate that refugees and sales influence the exchange rate, they do not provide evidence regarding the direction of the influence. We address this issue by examining the coefficients of the vector autoregressions and by considering the effect of unanticipated events on the exchange rate. In Table 3 we report the coefficients from the unconstrained regressions which include both refugees and sales, equations (3) and (4), from Table 1. The coefficients on the first lagged value of both refugees and sales are positive and significant, indicating that an increase in either causes the East mark to depreciate. This is consistent with the proposition regarding refugees, but not sales. The signs and significance levels of the other coefficients in the autoregressions are mixed.

Much recent work on exchange rate determination, motivated by the view of the exchange rate as an asset price which freely adjusts to reflect new information, focuses on the role of unanticipated movements in exogenous variables. The strongest expression of this view, the efficient markets hypothesis, states that no variable should Granger-cause the first difference of the ex-

TABLE 3—COEFFICIENTS FROM THE UNCONSTRAINED REGRESSION<sup>a</sup>  
Dependent Variable,  $E_t$ 

Lag	Independent Variables		
	<i>E</i>	<i>R</i>	<i>S</i>
(1)	-.07 (-.61)	.05 (2.60)	.12 (3.98)
(2)	.18 (1.60)	-.05 (-1.88)	-.11 (-3.84)
(3)	.00 (.01)	-.08 (-2.54)	.02 (.49)
(4)	.11 (1.06)	.04 (1.43)	.04 (1.18)
(5)	-.22 (-2.37)	-.03 (-1.15)	-.05 (-1.61)
(6)	-.23 (-2.56)	.04 (1.60)	-.01 (-.38)

<sup>a</sup>Unconstrained equation in (3) and (4) of Table 1. *t*-statistics are in parentheses. The constant of the regression,  $\alpha_0 = .30$  (1.29).

change rate. As seen in Table 1 this is obviously rejected for the exchange parlor rate. We therefore test for the influence of unanticipated movements in refugees and sales without assuming that the hypothesis holds.

Equation (1) reports the results of a regression with unanticipated movements in the exchange rate (EU) as the dependent variable and unanticipated movements in refugees (RU) and sales (SU) as the independent variables. EU is constructed by taking

the residuals from a regression of the exchange rate on lagged exchange rates, refugees, and sales (the unconstrained regression in line (3) of Table 1). RU and SU are the residuals from the same regression with R and S as the dependent variables. *t*-statistics are in parentheses.<sup>41</sup>

$$(1) \quad EU = .04RU + .03SU \quad R^2 = .08 \\ (2.63) \quad (1.14) \quad DW = 1.98$$

The effect of an unanticipated increase in refugees is of the expected sign, causing a depreciation of the East mark, and is statistically significant. This is consistent with the coefficients of the autoregression reported in Table 3. An unanticipated increase in retail sales does not have a significant effect on the exchange rate.

Refugees and sales were obviously not the only determinants of the exchange rate. In order to assess the influence of other economic and political factors, we include, in equation (2), dummy variables for several important events which occurred during the sample period. These are: D1 (October 1957)—a surprise currency swap in the GDR which rendered East mark balances in West Berlin worthless,<sup>42</sup> D2 (June 1958)—ration cards were abolished and prices raised for basic foodstuffs in East Germany,<sup>43</sup>

D3 (April 1960)—agricultural collectivization was completed across the GDR and D4 (October 1956)—the Hungarian revolution.

$$(2) \quad EU = .04RU + .01SU - .06D1 \\ (2.63) \quad (.59) \quad (-3.15) \\ + .05D2 + .04D3 + .03D4 \\ (2.94) \quad (1.95) \quad (1.79) \\ R^2 = .30 \\ DW = 1.90$$

The currency swap appreciated while the abolition of ration cards, agricultural collectivization, and the Hungarian revolution depreciated the East mark. Adding the four dummy variables does not change the significance of unanticipated refugees or the insignificance of unanticipated sales.<sup>44</sup>

## V. Conclusion

We have presented both anecdotal and statistical evidence relevant for the understanding of the market for the East mark in West Berlin over the period 1953–1961, the last years of complete freedom of movement between East and West Berlin. Applying conventional Granger causality tests to a monthly series of the West Berlin exchange parlor EM/DM exchange rate, we have found statistically significant evidence that both refugee flows and movements of retail sales in East Germany mattered for changes in the exchange rate. Unanticipated increases in the refugee flow were shown to lead to unexpected depreciations of the East German mark. Finally, important economic and political events were seen to have influenced the EM exchange rate as well.

## DATA APPENDIX

*Exchange Parlor Market Rate:* ( $E_t$ ).  $E_t = \ln(E_t^*) - \ln(E_{t-1}^*)$ , where  $E_t^*$  is end of month mean of the bid-and-ask price of 1 DM. Source: For August 2, 1948, through the end of December 1955, the exchange parlor

<sup>41</sup>The equation was estimated by OLS which, as shown by Adrian Pagan (1984), provides correct standard errors for unanticipated variables used as regressors in a two-step procedure.

<sup>42</sup>Exchange parlors were closed on October 14, 1957. Black-market traders exchanged between 3.5 to 7.5 EM/DM. Beginning October 22, East marks began to flow again significantly into West Berlin and various West Berlin department stores stopped accepting EM in payment and some exchange parlors limited the amount they would buy, with East mark depreciating to almost the same level as before the currency swap. Hans Reichardt et al., Vol. 8, 1974, pp. 282–3.

<sup>43</sup>The Committee for the East mark of the Berlin exchange parlors announced that only as many East marks would be bought over the weekend of May 31/June 1, 1958, as could be sold. Numerous East Berliners were expecting a currency swap to follow the ending of food ration cards. The ration cards were ended May 29 for meat and sausage, oils and fat, milk and sugar. Reichardt et al. (Vol. 8), p. 541.

<sup>44</sup>Including the four dummy variables in the vector autoregressions does not affect the results in Tables 1 and 2.

rate was reported on a daily basis in *Berlin in Zahlen*, 1950, 1951, and the *Statistisches Jahrbuch Berlin* (West), 1952–56. After the 1956 volume of the *Statistisches Jahrbuch Berlin* only monthly averages were published in that source. For the remaining sample, daily quotations were obtained from the West Berlin daily newspaper *Der Tagesspiegel*. We are grateful to Barbara Seitz and Lutz Spannagel for noting these quotations for us from microfilmed issues of *Der Tagesspiegel*.

*Refugee Flows:* ( $R_t$ ).  $R_t = \ln(R_t^*)$ , where  $R_t^*$  is the monthly total of refugees from East Germany registered at the emergency reception centers of West Berlin and the Federal Republic of Germany. *Source:* These data are taken from individual volumes of the *Statistisches Jahrbuch der Bundesrepublik Deutschland*. The published series begins September 1949.

*Retail Sales:* ( $S_t$ ).  $S_t = \ln(S_t^*)$ , where  $S_t^*$  are total retail sales in the GDR. *Source:* The series (in million EM) has been assembled from annual volumes of the *Statistisches Jahrbuch der Deutschen Demokratischen Republik*. The series begins January 1950.

## REFERENCES

- Baillie, Richard T., Lippens, Robert E. and McMahon, Patrick C., "Testing Rational Expectations and Efficiency in the Foreign Exchange Market," *Econometrica*, May 1983, 51, 553–63.
- Berlin in Zahlen, Berlin (West), 1950.
- Brabant, Jozef M. van, "Exchange Rates in Eastern Europe: Types, Derivation, and Application," *World Bank Staff Working Papers*, No. 778, Washington: The World Bank, 1985.
- Frenkel, Jacob A., "Exchange Rates, Prices, and Money: Lessons from the 1920's," *American Economic Review*, May 1980, 70, 235–42.
- Granger, Clive, "Investigating Causal Relations of Econometric Models and Cross-Spectral Methods," *Econometrica*, July 1969, 37, 424–38.
- Knetschke, Claus, "Der Handel D-Mark-West–D-Mark-Ost," unpublished Diplomarbeit, Leverkusen, FRG, 1956.
- Pagan, Adrian, "Econometric Issues in the Analysis of Regressions with Generated Regressors," *International Economic Review*, February 1984, 25, 221–47.
- Reichardt, Hans J., Drogmann, Joachim and Treutter, Hanns U., eds., *Berlin Chronik des Jahres*, Vol. 8, 1957–1958, Berlin (West): Heinz Spizing Verlag, 1974.
- Smith, Jean Edward, ed., *The Papers of General Lucius D. Clay: Germany 1945–1949*, 2 vols., Bloomington: Indiana University Press, 1974.
- Staatliche Zentralverwaltung für Statistik, *Statistisches Jahrbuch der Deutschen Demokratischen Republik*, Berlin (East): Staatsverlag der Deutschen Demokratischen Republik, 1955–1962.
- Statistisches Bundesamt, *Statistisches Jahrbuch für die Bundesrepublik Deutschland*, Stuttgart und Mainz (FRG): W. Kohlhammer GmbH, 1952–1962.
- Urban, Roswitha, "Die Wechselstuben," in Deutsches Institut für Wirtschaftsforschung, *Berlins Wirtschaft in der Blockade*, Berlin (West): Duncker & Humblot, 1949, 135–38.
- Wonnacott, Paul, *The Floating Canadian Dollar: Exchange Flexibility and Monetary Independence*, Washington: American Enterprise Institute, 1972.



# The Relative Efficiency of Slavery Revisited: A Translog Production Function Approach

By ELIZABETH B. FIELD\*

A debate over the relative efficiency of slavery has been ongoing for some time, sparked by the publication in 1974 of *Time on the Cross*, by Robert Fogel and Stanley Engerman. Fogel and Engerman conclude that "economies of scale in slaveholding made large slave farms more efficient than smaller ones, and more efficient than free farms" (Fogel and Engerman, 1974a, pp. 202-4). Although they identify a number of sources of scale economies, including such factors as indivisibilities in child care and in utilization of the labor of the elderly (Fogel and Engerman, 1974b, p. 141), they regard the use of the work gang system on large slave farms as the most important source of scale economies and the only one exclusively identified with the use of slave labor.

Fogel and Engerman (1974a) note that "there were no large-scale Southern farms based on free wage labor," and attribute this finding to nonpecuniary disadvantages of gang labor regimentation and the use of force (Fogel and Engerman, 1974a, p. 194).<sup>1</sup> Slave laborers, unlike free workers, could be compelled to work in gangs. Thus the efficiency advantage conferred by the specialization of labor and standardization of tasks associated with the gang system can be identified only with slave labor. Fogel and Engerman's argument suggests that slave and free labor may have been different, nonsubstitutable inputs at the margin.

Their hypothesis has intuitive appeal. Simply considering a description of the gang system, with its emphasis on interdependence and standardization of tasks, suggests that the system might indeed have led to greater efficiency. The regimentation and close supervision of the system might have led to greater work intensity. This paper investigates the arguments made by Fogel and Engerman (1980) regarding slave agriculture. Their hypothesis can be tested by estimating separate production functions for small and large slave farms, with free and slave labor disaggregated, and examining the estimated coefficients. The approach avoids some of the problems with Fogel and Engerman's original method, which have been noted by many critics.<sup>2</sup> The main approach in *Time on the Cross* is highly aggregative, dividing farms into four classes based on slaveholding size (Fogel and Engerman, 1974b, p. 139). The approach has many shortcomings. Donald Schaefer and Mark Schmitz (1979) have argued that Fogel and Engerman are simply observing ordinary economies of scale, such as those in child care, but that these economies could also have been achieved with free labor. To claim that slavery per se conferred an efficiency advantage, Fogel and Engerman would have had to distinguish between the two types of scale effects, which they cannot do with their aggregate approach.

In Volume II of *Time on the Cross: Evidence and Methods* (1974b), Fogel and Engerman present estimates of a Cobb-Douglas production function, which they use to distinguish "pure" from "incidental" scale effects (pp. 141-3), and in their 1980 paper, they again estimate a production function,

\*Department of Economics, Hamilton College, Clinton, NY 13323. This paper is a revision of part of my doctoral dissertation. I would like to thank Dudley Wallace, Robert Gallman, Gregg Lewis, Jean Mitchell, and the referee.

<sup>1</sup>Ralph Shlomowitz (1984) found gang system agriculture in the postbellum South in sugar cultivation, but not in cotton, following emancipation. He documents a great distaste for work under the gang system in the postbellum South and in several other countries.

<sup>2</sup>See Paul A. David et al., eds., 1976; David and Peter Temin, 1979; and Gavin Wright, 1979.

this time with a dummy variable that takes the value one for farms with over fifteen slaves.<sup>3</sup>

The approach is an improvement which avoids the problems of aggregation but is still not satisfactory. Although Fogel and Engerman (1974a,b, 1980) are implicitly stating that free and slave laborers differed as productive inputs, they employ an aggregate labor variable, and a functional form which makes strong assumptions about substitutability of factors.<sup>4</sup> This paper disaggregates free and slave labor, and employs a functional form placing no restrictions on substitutability. It will employ a separate production function for small and large farms. Fogel and Engerman (1980) use the same for both, although the gang system may well have represented a different production technology with different parameters. Finally, they make the division at fifteen slaves based on historical evidence, while this paper tests for the cutoff point.

### I. Data and Variables

The paper investigates cotton-growing slave farms in the late antebellum South. The data set used is the Parker-Gallman sample, drawn from the 1860 manuscript census.<sup>5</sup> Only slave farms are used.

<sup>3</sup>Fogel and Engerman (1980, pp. 686-7). They cite the historically derived threshold limit of fifteen slaves for use of the gang system, and contend that their method allows them to distinguish between use of the gang system and the other types of scale economies, which, they argue, were continuous unlike the gang system.

<sup>4</sup>It can be shown that the Cobb-Douglas form makes inappropriate separability restrictions for this sample. See my paper (1985, ch. 5) for a discussion.

<sup>5</sup>The data were made available by the Inter-University Consortium for Political and Social Research. The data for the 1860 Cotton Sample were originally collected by William N. Parker and Robert E. Gallman under a grant from the National Science Foundation. Neither the original collectors of the data nor the consortium bear any responsibility for the analyses or interpretations presented here. Fogel and Engerman's instructions for computing input and output variables were also used. The author will supply a copy of the data ( $Q, F, S, K, L$ ) as well as a listing of the program used to calculate these variables from the information

There were 2080 slave farms in the sample.<sup>6</sup> Of these, roughly 66 percent fell below Fogel and Engerman's threshold for employment of the gang system (fifteen or fewer slaves). These farms had an average annual output of \$969. Approximately 24 percent of the slaves in the sample lived on these farms, which contained about 40 percent of the total improved acreage of the farms sampled; their combined capital stock was 33 percent of the total in the slave farms sampled. About 28 percent of farms had between sixteen and fifty slaves. Such farms had an average output of \$3,695. Over 43 percent of slaves lived on these medium-sized farms, which represented 37 percent of the improved acreage and 37 percent of the capital stock. Fogel and Engerman's largest-size class, 51 slaves and over, contained only about 6 percent of the slave farms in the sample, but approximately 30 percent of the slaves. They contained 23 percent of the improved acreage and 30 percent of the total capital stock. Their average output was \$12,794.

The input and output definitions used in this analysis are largely those specified by Fogel and Engerman (1974b), but the output definition corrects double-counting of feed allowances. The output variable measures the value of net final output, with allowances for seed and livestock feed removed. The labor variables are expressed in prime-age male equivalents, following Fogel and Engerman, (1974b) but free and slave labor are disaggregated. Men and women are aggregated for both free and slave labor, and Fogel and Engerman's assumptions about free women's labor force participation rates are followed. The capital input is an annualized input, assuming a 10 percent rate of return. Finally, the land index consists of

on the ICPSR tape, and a description of the differences from Fogel and Engerman's procedure, on request. I relied on Frederick Bode and Donald Ginter, 1984; Richard Sutch, 1976; Robert Gallman, 1970; W. K. Hutchinson and S. Williamson, 1971; Martin Primack, 1977; and David and Temin, 1979, in making these modifications. For a detailed discussion see Field (1985, ch. 4).

<sup>6</sup>After exclusions, such as for nonpositive net output.

improved acreage multiplied by a soil quality index derived from information on soil types contained in the Parker-Gallman sample. This procedure was followed to avoid the inclusion of locational components, as would occur if land values were used.<sup>7</sup>

## II. The Model

Fogel and Engerman's hypothesis states that use of the gang system, a technology only employed with slave labor, led to greater efficiency on slave farms. Historical evidence leads them to conclude that this technology could only be employed past a threshold farm size, which they believe to have been about fifteen slaves.

To test this hypothesis, then, it must first be established that large farms employed a distinct production technology with different parameters from those of the small farm production function. Because the threshold of fifteen slaves is not directly observed, it is desirable to establish the most likely cutoff point by statistical methods. The coefficients of the two production functions must then be compared to establish whether this was a more efficient, or simply a different production technology. This can be tested in two ways. First, one can evaluate predicted output and the marginal product of slave labor for average input values at the threshold. If large farms were more efficient, predicted output and marginal product should be greater for the large farm production function. Because the other sources of scale economies are continuous, a jump in output would be evidence that the gang system was the source of the increase. It must also be established that, if  $f(x)$  is the small farm production function and  $g(x)$  the large farm

production function, and  $x^*$  the cutoff, that not only is  $g(x^*) > f(x^*)$  but that  $g(x) > f(x)$  for all  $x > x^*$ .

Previous production functions estimated for Southern agriculture have been the Cobb-Douglas and the CES varieties. Even if free and slave labor are disaggregated, both functional forms impose strong restrictions on the elasticities. Given that Fogel and Engerman's hypothesis (1974a,b, 1977, 1980) implies that free and slave labor differed as productive inputs, a flexible production function, not having any a priori restrictions on substitutability, would be preferable.

It is also desirable that the functional form chosen permit testing of the hypothesis of constant returns to scale, to check for economies of scale other than those associated with the gang system. In addition, linearity in parameters is desirable. The translog production function satisfies these requirements.

The empirical model is a physical production function which expresses the logarithm of output as a generalized quadratic function of the logarithms of the inputs.<sup>8</sup> Most em-

<sup>8</sup>The general form of the translog production function is

$$\begin{aligned} (1) \quad \ln Y = & \ln a_0 + a_A \ln A \\ & + \sum_i a_i \cdot (\ln X_i) + \frac{1}{2} b_{AA} (\ln A)^2 \\ & + \frac{1}{2} \sum_i \sum_j b_{ij} \cdot (\ln X_i) \cdot (\ln X_j) \\ & + \sum_i b_{iA} \cdot (\ln X_i) \cdot (\ln A), \end{aligned}$$

where  $Y$  is output, the  $X_i$  are factors of production,  $A$  is an index of technology, and  $b_{ij} = b_{ji}$ . Assuming homogeneity implies that

$$\begin{aligned} (2) \quad \sum_i b_{ij} &= 0 = \sum_j b_{ij}, \\ \sum_i \sum_j b_{ij} &= 0, \sum_i b_{iA} = 0. \end{aligned}$$

For linear homogeneity, or constant returns to scale, in addition,  $\sum_i a_i = 1$ .

Hicks-neutral technical change implies  $a_A = 1$ ,  $b_{AA} = 0$ , and  $b_{iA} = 0$ , so that the production function be-

<sup>7</sup>Fogel and Engerman (1977, 1980) and David and Temin (1979) disagree about the calculation of the land input. Fogel and Engerman remove the estimated value of the locational component of land values, while David and Temin inflate the locational component for all sizes of farm to that of the largest farms. I have chosen to use to approach the problem more directly. I use improved acreage, not the value of that acreage (which avoids the problem of locational components altogether) and weight the acreage by a soil quality index, to reflect differences in natural fertility of the land.

pirical work using quadratic production functions assumes constant returns to scale, so that the logarithmic marginal products can be interpreted as cost shares, and cost share equations estimated. But constant returns to scale is not necessarily an appropriate assumption for slave farms, and consequently the physical production function will be directly estimated using ordinary least squares.

### III. Empirical Results

A production function was first estimated for all slave farms and then for those farms with over fifteen slaves and those at or under that number, following Fogel and Engerman (1974b). A Chow test was performed, and the hypothesis that the production parameters were the same for the two size classes was rejected. The large farms were further split at fifty slaves, again following Fogel and Engerman. However, in this case, the hypothesis that farms over fifty slaves had the same production function could not be rejected.<sup>9</sup> Thus it appears that large farms (sixteen slaves and over) did have a different production technology, but that there was no further division.

Fogel and Engerman base the split at fifteen slaves on examination of managerial

documents of slave plantations. However, it is desirable to confirm this statistically. By splitting the data set at various values of the slaveholding variable and estimating two translog functions for each cutoff point (for those observations lying above and below the split), one can determine the maximum likelihood estimate of the value of the cutoff. The decision was made to use the adjusted slave labor force (prime-age male equivalents) rather than the number of slaves held as a switching variable, since it seems likely that the number of adult male equivalent slaves was important in determining the adoption of the gang system, and the adjusted slave labor force weights these individuals more heavily. Another possible switching variable would be the number of prime-age males, which would avoid the situation of no adult males, but many women and children, which can also occur if the number of slaves is used. However, it is not desirable to exclude women and children entirely, although they should be weighted less; the switching variable used will be  $S$ .

The cutoff chosen was a value of  $S$ , the adjusted slave labor variable, of six.<sup>10</sup> This corresponds to approximately fifteen slaves held.<sup>11</sup> Thus the statistical procedure and Fogel and Engerman's historically based cutoff (1974a, 1977, 1980) give the same result: there was a switch in regimes, with the switching point at about fifteen slaves. This is very suggestive. Although there is no direct evidence that the switch was due to institution of the gang system, the fact that the statistically chosen switch corresponds to historical evidence about the threshold for use of the gang system strongly implies that

comes

$$(3) \quad \ln Y = \ln A + \ln a_0 + \sum_i a_i \cdot (\ln X_i) \\ + \frac{1}{2} \sum_i \sum_j b_{ij} \cdot (\ln X_i) \cdot (\ln X_j).$$

Constant returns to scale may not be an appropriate assumption for the antebellum South, since economies of scale are hypothesized. However, it can easily be shown that, with a slight relaxation of the adding up constraints, the translog function with symmetry and the new adding-up conditions imposed is a quadratic approximation to an arbitrary homogeneous production function.

<sup>9</sup>For the split at 15 slaves, the test statistic  $F$  had a value of 4.38 with 15 and 2050 d.f. and the null hypothesis of no switch was rejected at a significance level of .05. For a further split at 50, the test statistic had a value of .603 with 15 and 676 d.f.; the null hypothesis could not be rejected.

<sup>10</sup>Because there is no direct evidence on where the cutoff point occurred, a method for estimation of the parameters of a linear system with parameters obeying two regimes, with slave labor as the separation variable, will be used. See Richard Quandt (1958) for a description of the likelihood function. Note that noninteger values of  $S$  were not tested, so that it is possible that the switch actually occurred at an intermediate value of  $S$ . The values of  $S$  tested were 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 15.

<sup>11</sup>Of twenty-five farms with values of  $S$  between 5.9 and 6.1, the average value of slaves held was 14.92.

the switch was due to implementation of the gang technology.

Each of the two production functions ( $S \leq 6$  and  $S > 6$ ) was then tested for homogeneity and linear homogeneity (constant returns to scale). The test for constant returns to scale (CRS) is particularly interesting because economies of scale besides the use of the gang system have been suggested, such as housing, child rearing, and use of the labor of the elderly. None of these seems likely to have had a definite threshold value and can thus be examined by testing for CRS. Homogeneity and linear homogeneity are accepted on both large and small farms, however. Although there were slightly increasing returns to scale in both categories, the degree of homogeneity was not significantly different from one.<sup>12</sup> The estimates of the parameters with CRS imposed are presented in Table 1 and form the basis for the tests that follow.

Although other economies of scale may have been present, they were not significant, and thus the remaining source, use of the gang system, is examined. First, the two production functions are evaluated using average input values for the cutoff. If the gang system did confer greater efficiency, one would expect a jump in output and the marginal product of slave labor. Using the small farm production function, predicted output was \$1422 and the estimated value of marginal product for slave labor was \$97. For the large farm function, predicted output was \$1689 and the value of the marginal product of slave labor was \$119. Thus, both output and marginal product jump at the threshold.<sup>13</sup>

<sup>12</sup> The estimated degree of homogeneity was 1.07 on small farms; the test statistic  $F$  had a value of 2.52 with 1, 1341 d.f. The estimated degree of homogeneity on large farms was 1.11; the test statistic had a value of 3.65 with 1, 709 d.f. In both cases, CRS could not be rejected.

<sup>13</sup> Treating average predicted output and value of marginal product of slave labor for large and small farms at the threshold as independent random variables, the test statistic  $t$  for differences in predicted output was 5.878; that for differences in predicted value of marginal product was 5.0417.

TABLE 1—THE DETERMINANTS OF THE LOGARITHM OF THE MARKET VALUE OF OUTPUT ( $Q$ )

The Coefficient of:	Small Farms	Large Farms
Constant	.339 (.314)	-.913 (.972)
Log of Free Labor ( $\ln F$ )	-.548 (.236)	-.444 (.248)
Log of Slave Labor ( $\ln S$ )	-.317 (.270)	-.801 (.402)
Log of Capital ( $\ln K$ )	1.431 (.234)	1.327 (.303)
Log of Land ( $\ln L$ )	.435 (.253)	.918 (.345)
$(\ln F)^2$	-.001 (.052)	-.094 (.046)
$(\ln S)^2$	.031 (.068)	-.358 (.133)
$(\ln K)^2$	-.359 (.061)	-.214 (.086)
$(\ln L)^2$	-.076 (.031)	-.080 (.093)
$(\ln F) \cdot (\ln S)$	-.161 (.047)	.052 (.063)
$(\ln F) \cdot (\ln K)$	.219 (.046)	-.005 (.048)
$(\ln F) \cdot (\ln L)$	-.057 (.046)	.048 (.053)
$(\ln S) \cdot (\ln K)$	.072 (.054)	.247 (.084)
$(\ln S) \cdot (\ln L)$	.054 (.055)	.059 (.085)
$(\ln K) \cdot (\ln L)$	.068 (.042)	-.028 (.083)
Adjusted $R^2$	.46	.65
$n$	1356	724

Note: Standard errors are given in parentheses.  $Q$  = market value of final output less feed and seed allowances.  $F$  = prime-age male equivalent free labor.  $S$  = prime-age male equivalent slave labor.  $K$  = annualized capital input.  $L$  = improved acreage times soil quality index.

It must also be established that the gang system conferred an advantage, not simply at the threshold, but for all farms above the threshold. Predicted outputs using the small farm production parameters and using the large farm parameters were compared for all farms with sixteen or more slaves. At 84 percent of the data points, predicted output using the large farm production function was higher. It seemed likely that the other 16 percent lay close to the threshold, and perhaps had been incorrectly classified as gang farms. However, that did not prove to be the case. For the 16 percent of large farms where

the predicted difference was negative, the number of slaves held was actually higher on average. Actual output was lower, and the land and capital variables slightly smaller than in that part of the sample with a positive difference. These farms appear to have been less efficient.

Although the large farm production function does not give larger predicted outputs for all data points, it does for the majority. Furthermore, those farms where predicted outputs were larger using the small farm coefficients appear to have differed from the average large farm in having a larger slave labor force, lower output, and lower capital to labor and land to labor ratios. In addition, output and the marginal product of slave labor jump significantly at the threshold. Therefore, large slave farms were more efficient than small farms.

Fogel and Engerman (1980, p. 675) found an efficiency advantage of approximately 38 percent to the gang system. On the large farms where the predicted difference was positive, this paper finds an average advantage of approximately 31 percent, while on all large farms, the average advantage was about 24 percent.

#### IV. Conclusion

It has been argued that large slave farms were more efficient than small slave farms and free farms (Fogel and Engerman, 1974a) and that this superior efficiency derived from the work gang system found only on large slave farms. Historical evidence shows that the system was used with sixteen or more slaves. This paper tests Fogel and Engerman's (1974a,b, 1977, 1980) hypothesis concerning slave agriculture, using methods that avoid problems of earlier analyses: The empirical results show that a switch in regimes between small and large farms did occur at about fifteen slaves. Although there were no significant economies of scale in the ordinary sense under either regime, the switch from the small to the large farm regime did permit a significant increase in output and caused the marginal slave to be more productive. Furthermore, the large farm production function predicted higher output on the

majority of large farms. Large farms were 24 percent more efficient on average. The fact that the statistically chosen switching point coincided with the threshold for use of the gang system found in historical documents by Fogel and Engerman suggests strongly that the increase in efficiency was due to the use of the gang system.

The increase in efficiency is associated only with the use of slaves. The gang system was not used with free labor in the antebellum South. This may have been because prevailing wages were insufficient to induce free workers to endure the adversities of gang work, or it may have been due to other factors such as the difficulty of guaranteeing an adequate supply of free labor on the farm during the planting and harvest seasons. Whatever the reason, the increase in efficiency in cotton production associated with use of the gang system was only captured by the use of slaves.

#### REFERENCES

- Bode, Frederick A. and Ginter, Donald E., "A Critique of Landholding Variables in the 1860 Census and the Parker-Gallman Sample," *Journal of Interdisciplinary History*, Autumn 1984, 15, 277-95.
- Christensen, Laurits R., Jorgenson, Dale W. and Lau, Lawrence J., "Conjugate Duality and the Transcendental Logarithmic Production Function," *Econometrica*, July 1971, 39, 225-56.
- \_\_\_\_\_, \_\_\_\_\_ and \_\_\_\_\_, "Transcendental Logarithmic Production Frontiers, *Review of Economics and Statistics*, February 1973, 55, 28-45.
- David, Paul A. et al., eds., *Reckoning with Slavery: A Critical Study in the Quantitative History of American Slavery*, New York: Oxford University Press, 1976.
- \_\_\_\_\_, and Temin, Peter, "Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South: A Comment," *American Economic Review*, March 1979, 69, 213-8.
- Field, Elizabeth, "Elasticities of Complementarity and Returns to Scale in Antebellum Cotton Agriculture," unpublished doctoral

- dissertation, Duke University, 1985.
- Fogel, Robert W. and Engerman, Stanley L., (1974a) *Time on the Cross: The Economics of American Negro Slavery*, Boston: Little, Brown, 1974.
- \_\_\_\_\_ and \_\_\_\_\_, (1974b) *Time on the Cross: Evidence and Methods — A Supplement*, Boston: Little, Brown, 1974.
- \_\_\_\_\_ and \_\_\_\_\_, "Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South," *American Economic Review*, June 1977, 67, 275-96.
- \_\_\_\_\_ and \_\_\_\_\_, "Explaining the Relative Efficiency of Slave Agriculture in the Antebellum South: Reply," *American Economic Review*, September 1980, 70, 672-90.
- Gallman, Robert E., "Self-Sufficiency in the Cotton Economy of the Antebellum South," *Agricultural History*, January 1970, 44, 5-23.
- Hutchinson, W. K. and Williamson, S., "The Self-Sufficiency of the Antebellum South: Estimates of the Food Supply," *Journal of Economic History*, September 1971, 31, 591-612.
- Primack, Martin Leonard, *Farm Formed Capital in American Agriculture: 1850 to 1910*, New York: Arno Press, 1977.
- Quandt, Richard E., "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes," *Journal of the American Statistical Association*, June 1958, 53, 324-30.
- Schaefer, Donald F. and Schmitz, Mark D., "The Relative Efficiency of Slave Agriculture: A Comment," *American Economic Review*, March 1979, 69, 208-12.
- Shlomowitz, Ralph, "Plantations and Smallholdings: Comparative Perspectives from the World Cotton and Sugar Cane Economies, 1865-1939," *Agricultural History*, January 1984, 58, 1-16.
- Sutch, Richard, "The Care and Feeding of Slaves," in Paul David et al., eds, *Reckoning with Slavery*, New York: Oxford University Press, 1976, 231-301.
- Wright, Gavin, "The Efficiency of Slavery: Another Interpretation," *American Economic Review*, March 1979, 69, 219-26.

# The Private *R&D* Investment Response to Federal Design and Technical Competitions

By FRANK R. LICHTENBERG\*

It is the federal government's role to promote investment in research and development (*R&D*) that yields innovations in the production of public goods, such as armaments for national defense and equipment for space exploration. At first blush, it might appear that the government can bring about such *R&D* investment by a combination of only two institutional arrangements: (1) performing *R&D* in government laboratories; and (2) contracting with private firms and nonprofit organizations (such as universities) to perform *R&D*. Indeed, official estimates of the distribution of U.S. *R&D* by mission (defense versus civilian) are based on the assumption that only *R&D* conducted under these two arrangements is directed toward federal missions.<sup>1</sup>

The major thesis to be developed in this paper, however, is that there is a third, and quantitatively important, method by which the government promotes *R&D* investment relevant to the provision of social goods: awarding major contracts by a method of acquisition known as "procurement by design and technical competition." In essence, this method consists in the government's simply revealing its demand for certain types of technological innovations, and encouraging private firms to sponsor the necessary *R&D*, the costs of which the sponsor will recover from profits on the sale of the product. Edwin Mansfield (1971) has observed

that before World War II, the government tended to issue relatively few *R&D* contracts to industry or universities. If the government wanted private firms to perform defense-related *R&D*, it would encourage them to finance it on their own.<sup>2</sup> Since the beginning of World War II, the real value of government *R&D* contracts has increased very rapidly, and contracting has apparently become the most important of the three methods used by the government to promote *R&D* investment related to social goods. But we shall provide evidence that the government continues to induce a considerable amount of privately financed, federal mission-oriented *R&D* expenditure by sponsoring design competitions.

We develop this evidence by estimating regressions of company-sponsored (private) *R&D* expenditure on the value of competitive and noncompetitive government contracts and on other (nongovernment) revenue. These regressions are estimated using annual firm-level panel data for the years 1979–84 (which spanned a major defense buildup) for a sample of 169 industrial companies. The coefficients of these equations are estimates of the marginal response of private *R&D* to changes in the volume of competitive and noncompetitive procurement and in nongovernment demand. We also distinguish between the effects on private *R&D* of government procurement of *R&D* and of non-*R&D* services and products.

Previous studies have suggested that federal (contract) *R&D* has a smaller (often

\*Columbia University Graduate School of Business, 726 Uris Hall, New York, NY, 10027, and National Bureau of Economic Research. Financial support from the National Science Foundation (via grant PRA 85-12979 to NBER), and from the MacArthur Foundation Faculty Research Program at Columbia University, is gratefully acknowledged. Two anonymous referees provided helpful comments on an earlier draft. Donald Siegel provided capable research assistance. I am responsible for any errors.

<sup>1</sup>See National Science Board, 1985, p. 6.

<sup>2</sup>"For example, in the procurement of military aircraft, an open competition was held; the winning firm recovered its development costs in the form of profits on the sale of the airplanes, and the losing firms did not recoup their *R&D* investment" (Mansfield, 1971, p. 122).



insignificant) effect on various measures of economic performance such as productivity and profitability than privately financed R&D.<sup>3</sup> Because the social returns to R&D done in pursuit of the federal government's objectives may differ from the social returns to other R&D investment, measuring the extent to which private R&D is dedicated toward federal missions may lead to a better understanding of the contribution of R&D to U.S. economic growth.

The remainder of this paper is organized as follows. In Section I we briefly describe the design and technical competition method of contracting, consider why the government chooses to employ this method (instead of or in addition to R&D contracting), and review previous evidence on private, federal mission-oriented R&D. In Section II we discuss the specification and estimation of an econometric model for determining the private R&D investment response to competitive procurement and to other components of demand. Empirical results are reported in Section III, and Section IV contains a summary and concluding remarks.

### I. Federal Design and Technical Competitions

A design and technical competition (henceforth abbreviated "design competition") begins "officially" when a federal agency (such as the Department of Defense) issues a formal Request for Proposals, which is often 1100 to 2500 pages long.<sup>4</sup> Three or four firms typically submit proposals (which may range from 23,000 to 38,000 pages in length) in response to requests issued by the Defense Department, which then begins an elaborate review process. The firm that submits the proposal receiving the highest "score" is

generally selected as the contractor, and is essentially guaranteed (unless Congress decides to cancel the project) to be awarded a sequence of contracts for R&D, production, spare parts, maintenance, training, and so forth, over a number of years. The contracts that are initially awarded to the successful firm are officially designated as "competitive" contracts. But most of the revenue that the firm will receive by virtue of having won the competition will come from subsequent, "follow-on" contracts, which are officially designated as "noncompetitive" contracts. Table 1 indicates that in fiscal year 1984, the value of noncompetitive follow-on contracts after design competition was 2.72 times as large as the value of competitive contracts associated with these competitions. Because the winner of the design competition is virtually assured of eventually being awarded the relatively large follow-on contracts, it is often suggested that contractors are willing to incur losses (by "buying in," or submitting bids below anticipated costs) on the initial competitive contracts.

I think that it is natural to pose the question of *why* the government does promote mission-oriented R&D by sponsoring design competitions, in addition to doing so by directly contracting with firms to perform R&D. Perhaps the most compelling possible explanations are based on imperfect information considerations. A design competition appears to be an almost perfect example of a *contest*, in the sense defined by Barry Nalebuff and Joseph Stiglitz (1983) and others. In a contest, or competitive compensation scheme, the individual's reward or compensation (for example, whether he is awarded a contract) is determined only by his *rank* vis-à-vis his competitors, rather than by his "individualistic" output (or marginal product), as is the case in the classical model of pure price competition. These authors have shown that when the principal (in our case, the government) cannot directly and costlessly observe the level of input (effort) of the agents (contractors), rewards based on relative output are superior to payments based on individualistic output. They argue that competitive compensation schemes have "...greater flexibility and greater adaptabil-

<sup>3</sup>See, for example, Zvi Griliches and Frank Lichtenberg, 1984, and Griliches, 1986.

<sup>4</sup>Many analysts have noted, however, that potential contractors are aware of the government's interest in particular areas of technology, and attempt to influence the "shape" of government demand for innovations, long before the publication of Requests for Proposals. See, for example, Charles Danhof, 1968.

TABLE 1—DISTRIBUTION OF DEPARTMENT OF DEFENSE FISCAL YEAR 1984  
NEGOTIATED COMPETITIVE AND NONCOMPETITIVE PROCUREMENT, BY METHOD

Method of Procurement	All contracts	R&D contracts <sup>a</sup>
<b>Competitive</b>		
Design and technical competition	11.6	4.4
Price competition	35.0	0.4
<b>Noncompetitive</b>		
Follow-on after design and technical competition	31.6	4.6
Follow-on after price competition	4.1	0.1
Catalog or market price	0.9	b
Other noncompetitive	34.3	3.9
<b>Total, all methods</b>	<b>117.6</b>	<b>13.4</b>

Source: Department of Defense (1985). All figures in billions of dollars.

<sup>a</sup>Contracts citing statutory authority 10 U.S.C. 2304(4)(11) ("experimental, developmental, test, or research") as their authorization for exception to the requirement for formal advertising.

<sup>b</sup>Less than \$0.05 billion.

ity to change in the environment than do other forms of compensation" (p. 41), so that contests may be preferred when the risk associated with common environmental variables (for example, the difficulty of achieving technical progress in a given area) is large. Moreover, "the use of a contest as an incentive device can induce agents to abandon their natural risk aversion and adopt 'riskier' and more profitable production techniques" (p. 23).

Inability of the principal to monitor the (relative) ability, or productivity, of various agents, is a second type of imperfect information which may render competitive compensation schemes optimal. Suppose that both government contracts and potential contractors are heterogeneous, in the sense that one firm is more qualified to perform a given contract than other firms, but that the government is uncertain about the identity of the most qualified supplier. A number of theoretical models of markets characterized by this kind of imperfect information show that it is equilibrium behavior for sellers to invest in acquiring, and for buyers to rely on, *signals* of quality and ability.<sup>5</sup> It may be

useful to interpret design competitions as signaling phenomena, in which the signal of contractor ability that the government relies on is the score on the technical proposal.

Factors other than imperfect information about contractor effort and/or ability may account for the existence of design competitions. For example, judging from periodic congressional hearings and reports on the subject (see, for example, U.S. Congress, 1969), there is strong congressional demand for competition in procurement, perhaps because Congress believes that competitive procurement promotes economic efficiency or fairness. The recent passage of the Competition in Contracting Act also reflects this demand.

Existing evidence concerning private investment in federal mission-oriented R&D is limited and fragmentary; we are not aware of any previous *econometric* evidence. Aside from case studies and anecdotal evidence concerning company proposal efforts related to specific design competitions,<sup>6</sup> we are familiar with only two pieces of evidence. The first consists of financial data on Independent Research and Development published by the Defense Department.<sup>7</sup> Inde-

<sup>5</sup>See, for example, A. Michael Spence, 1984; and Richard Kihlstrom and Michael Riordan, 1984.

<sup>6</sup>See, for example, Edward Roberts, 1969.

<sup>7</sup>For a detailed discussion of Independent R&D, see my paper, 1986, and Judith Reppy, 1977.

pendent *R&D* is contractor-initiated and contractor-directed technical effort that is not sponsored by, or required in performance of, a contract or grant. The Defense Department recognizes independent *R&D* as a necessary cost of doing business and reimburses contractors for a fraction (on average, about 40 percent) of independent *R&D* costs. But the Defense Department and its contractors regard *all* independent *R&D* expense as "company-funded" *R&D*, and contractors generally report it as such in the *National Science Foundation Survey*, which is the basis for the official estimates of industrial *R&D*. The Defense Procurement Regulations specify that in order for the costs of a project to be eligible for reimbursement, the project must have a "potential military relationship" to Defense Department functions and operations. In 1983, major defense contractors reported having incurred \$3.9 billion in independent *R&D* costs; this represents 9.2 percent of the National Science Foundation's estimate of \$42.6 billion for "company-funded" *R&D* in that year.

The second piece of evidence is provided by F. M. Scherer's (1984) analysis of "linked" *R&D* and patent data of the largest *R&D*-performing companies. Scherer attempted to classify each of about 15,000 U.S. patents (obtained by 443 companies between June 1976 and March 1977) by "industry of use," that is, to identify the sector(s) of the economy in which (most intensive) use of the invention was anticipated. Two of the industries of use defined by Scherer were "defense and space operations" and "government, except postal and defense." He estimated the value of company-sponsored *R&D* "used" by these sectors to be \$1206.3 million and \$378.7 million, 11.3 and 3.6 percent, respectively, of the total amount of company-funded *R&D* (\$10.64 billion) attributed to these companies.<sup>8</sup> Thus, according to Scherer's methodology, the federal

government is the primary beneficiary of about 15 percent of company-sponsored industrial *R&D* expenditure.

## II. Econometric Specification and Estimation Issues

Our research design for measuring the private *R&D* investment response to government procurement in general, and design competitions in particular, is to estimate, using longitudinal firm-level data, regressions of private *R&D* expenditure on three variables: the value of the firm's competitive contracts, of its noncompetitive contracts, and of its nongovernmental sales. (The sum of these three variables is total sales.) The coefficient of the first variable is our primary concern; we include the other two mainly to "control" for their influence and to provide benchmarks against which to measure the effect of competitive procurement.

As Table 1 reveals, there are two types of competitive procurement: design and technical competition, and price competition. Unfortunately, we cannot distinguish at the firm level between the two types of competitive contracts; we observe only the *sum* of the two.<sup>9</sup> Because price-competitive procurement is hypothesized to elicit less private *R&D* than design-competitive procurement, the coefficient on competitive contracts may perhaps be regarded as a lower-bound estimate of the effect of design-competitive procurement.

For several reasons—omitted variables, simultaneity, and errors in variables—the right-hand side variables may be correlated with the disturbance of the regression described above, so that ordinary least squares estimates may not be consistent. If omitted variables were the only problem, *and* if those variables were constant over *t* for a given *i*, then we could obtain consistent estimates by including a dummy variable for each firm,

<sup>8</sup>These estimates were obtained using what Scherer termed the "private goods" assumption, according to which a patent (and its associated *R&D* expenditure) benefited (was assigned to) only one, rather than several, industries of use.

<sup>9</sup>We can and do distinguish, however, between *R&D* and non-*R&D* competitive contracts. This is helpful because most competitive *R&D* contracts are awarded by design competition, whereas most competitive non-*R&D* contracts are awarded by price competition.

TABLE 2—SAMPLE AGGREGATES AND COMPARISON NATIONAL DATA

Year	Aggregates for Sample of 169 Firms			
	Company-Sponsored R&D Expenditure	Sales	Value of Federal Contracts	Value of Federal R&D Contracts
1979	17.4	732.8	38.3	7.0
1980	20.1	807.3	43.0	7.8
1981	23.0	887.5	52.6	8.3
1982	25.6	877.6	66.6	12.1
1983	28.0	908.6	72.7	13.5
1984	33.3	968.7	78.9	14.8

Year <sup>a</sup>	U.S. National Data			
	"Company and Other Funds for R&D" <sup>b</sup>	Sales <sup>c</sup>	DoD Prime Contract Awards <sup>c</sup>	"Federal Funds for R&D" <sup>b</sup>
1979	25.7	1215.0	58.5	12.5
1980	30.5	1427.6	66.7	14.0
1981	35.4	1589.2	87.2	16.4
1982	39.5	1479.3	102.5	18.5
1983	42.6	1596.5	121.1	20.2
1984	47.3 <sup>e</sup>	<sup>f</sup>	124.9	22.0 <sup>e</sup>

<sup>a</sup> All data except Defense Department contracts on calendar-year basis; contract awards on a fiscal-year basis. All amounts in billions of dollars.

<sup>b</sup> Source: National Science Board, 1985, Appendix Table 4-4.

<sup>c</sup> Net sales of R&D-performing manufacturing companies. (Source: National Science Board, 1985, Appendix Table 4-7.)

<sup>d</sup> Department of Defense Prime Contract Awards to Businesses for Work in the U.S. (Source: Department of Defense, 1985, Table 3).

<sup>e</sup> Estimated.

<sup>f</sup> Not available.

that is, by using a fixed-effects or "within" estimator. But the omitted variables may not be time-invariant, the within estimator does not eliminate possible simultaneous-equations bias, and it is likely to exacerbate biases due to errors in variables.<sup>10</sup> Fortunately, however, under reasonable assumptions we can address all three of these potential specification errors by estimating the model using instrumental variables. Below we report both ordinary least squares and instrumental variables estimates of "total" (excluding fixed-firm effects) and "within" versions of the model. We should perhaps pay the most attention to the instrumental variables total estimates; using instrumental

variables with fixed effects is superior only if the instruments are endogenous with respect to the omitted time-invariant characteristics, a possibility which seems unlikely to pose a serious problem.<sup>11</sup>

Because the model we estimate is a relationship among *levels* (rather than logarithms or ratios) of variables, the disturbances are likely to be heteroskedastic. In order to eliminate the heteroskedasticity we performed *weighted* least squares and instrumental variables estimation, using the reciprocal of total sales as the weight. We tested for homoskedasticity using the statistic proposed by Halbert White (1980); in all

<sup>10</sup> See Zvi Griliches and Jerry Hausman, 1986.

<sup>11</sup> The instruments used for the regressors of the model are described in the Data Appendix.

cases the  $\chi^2$  statistic was much lower than the critical value for rejecting homoskedasticity.

### III. Empirical Results

Variants of the model were estimated on annual 1979–84 panel data for 169 industrial firms. The construction of the data base is described in the Data Appendix. Sample aggregate values of selected variables and comparison national data are presented in Table 2. The data reflect the major defense buildup which occurred during this period: the value of federal contracts more than doubled, whereas total sales increased by only about 35 percent.

The estimates are presented in Table 3. Panel A of the table displays estimates of the model in which the coefficients on competitive and noncompetitive procurement are constrained to be equal. This model is useful for developing an estimate of the fraction of the total “induced” increase in private *R&D* accounted for by the increase in government procurement. Because the instrumental variable total estimates are most likely to be consistent and efficient, we focus mainly on these estimates. These imply that a \$1 increase in government sales tends to induce a 9.3-cent increase in private *R&D* while a \$1 increase in nongovernment sales induces only a 1.7-cent private *R&D* increase. The difference between these two coefficients is highly significant:  $F_{1,1006} = 31.6$  ( $P$ -value = .0001).<sup>12</sup>

These estimates enable us to compute the fraction  $\pi$  of the total induced increase in private *R&D* induced by the increase in government procurement.<sup>13</sup> The point estimate (standard error) of  $\pi$  is .528 (.050). The point estimate implies that slightly over half of the total induced increase in private *R&D*

between 1979 and 1984 was induced by the increase in government sales; the limits of the .95 confidence interval on this share are .430 and .626.

Panel B of Table 3 presents unconstrained estimates of the model. The difference between the coefficients on competitive and noncompetitive contracts is positive and highly significantly different from zero for all estimation techniques except ordinary least squares total, in which case it is negative and insignificant. The magnitude of the difference increases as we move across the columns from left to right. The total instrumental variables estimates suggest that a \$1 increase in competitive procurement induces a 54-cent increase in private *R&D* expenditure. As noted earlier, the prospect of substantial future noncompetitive follow-on contracts awarded to the winner of the design competition makes firms willing to make *R&D* investments which are large, relative to the value of the initial competitive contracts. The coefficient on noncompetitive contracts is negative but insignificant, suggesting that the entire stimulus to private *R&D* from government procurement comes from competitive acquisition.

The remainder of the empirical analysis is devoted to extending and amplifying the major finding that a considerable quantity (and share) of private *R&D* investment is induced by competitive procurement. The *timing* of private *R&D* investment, relative to the award of contracts, is the issue we address first. The regressions reported above are of company-funded *R&D* on *contemporaneous* sales classified by type. It is natural to hypothesize, however, that some private *R&D* expenditure is made by firms in years prior to the year in which competitive contracts are awarded. To investigate this hypothesis, we estimated the model including also values of the regressors in year  $t+1$ ; these estimates are shown in panel C. In all four columns, the coefficient on future competitive contracts is almost as large as or larger than the coefficient on current competitive contracts; the future coefficient is always more significant. The instrumental variables total estimates imply that most of the long-run response of private *R&D* to

<sup>12</sup>In the case of the instrumental variable within estimates, the difference is not significant:  $F = 0.3$ .

<sup>13</sup> $\pi = \gamma_1 \Delta G / (\gamma_1 \Delta G + \gamma_2 \Delta N)$ , where  $\Delta G$  denotes the change in government sales,  $\Delta N$  denotes the change in nongovernment sales, and  $\gamma_1$  and  $\gamma_2$  are the coefficients on  $G$  and  $N$ , respectively. As shown in Table 2, the sample aggregate values of  $\Delta G$  and  $\Delta N$  over the period 1979–84 were \$40.6 and \$195.3 billion, respectively.

TABLE 3—ORDINARY LEAST SQUARES AND INSTRUMENTAL VARIABLES ESTIMATES OF WEIGHTED TOTAL AND WITHIN REGRESSIONS OF COMPANY-FUNDED R&D ON SALES CLASSIFIED BY TYPE  
(*t*-Statistics in Parentheses)

Independent Variables	Ordinary Least Squares		Instrumental Variables	
	Total	Within	Total	Within
A				
Government Contracts	0.046 (10.3)	0.050 (9.95)	0.093 (7.12)	0.027 (1.08)
Nongovernment Sales	0.027 (33.4)	0.034 (20.4)	0.017 (7.46)	0.041 (6.38)
B				
Competitive Contracts	0.039 (1.79)	0.105 (6.38)	0.544 (4.85)	0.694 (2.57)
Noncompetitive Contracts	0.048 (6.53)	0.041 (7.12)	-0.040 (1.06)	-0.153 (1.74)
Nongovernment Sales	0.027 (33.4)	0.034 (20.2)	0.017 (6.35)	0.038 (3.31)
C				
Competitive Contracts ( <i>t</i> )	-0.044 (0.96)	0.062 (3.23)	0.085 (0.35)	0.300 (1.05)
Competitive Contracts ( <i>t</i> + 1)	0.072 (1.77)	0.056 (3.56)	0.400 (1.85)	0.498 (1.55)
Noncompetitive Contracts ( <i>t</i> )	0.033 (1.42)	0.031 (3.79)	0.097 (0.84)	-0.012 (0.11)
Noncompetitive Contracts ( <i>t</i> + 1)	0.015 (0.73)	0.007 (0.93)	-0.116 (1.13)	-0.193 (1.14)
Nongovernment Sales ( <i>t</i> )	0.002 (0.35)	0.022 (11.0)	-0.026 (0.87)	-0.001 (0.03)
Nongovernment Sales ( <i>t</i> + 1)	0.022 (4.33)	0.009 (4.67)	0.040 (1.43)	0.044 (1.82)
D				
Competitive R&D	-0.048 (1.29)	0.085 (1.87)	0.857 (1.01)	0.174 (0.08)
Noncompetitive R&D	0.156 (3.09)	0.005 (0.17)	-2.113 (2.18)	-1.683 (1.24)
Competitive non-R&D	0.072 (2.07)	0.123 (6.05)	1.212 (3.89)	1.077 (1.80)
Noncompetitive non-R&D	0.036 (4.05)	0.046 (6.42)	-0.074 (0.96)	-0.050 (0.32)
Nongovernment Sales	0.027 (33.5)	0.034 (20.2)	0.016 (3.53)	0.037 (2.03)
E				
R&D Contracts	0.039 (1.98)	0.047 (2.56)	-0.476 (2.63)	-0.930 (2.05)
Non-R&D Contracts	0.048 (7.09)	0.051 (7.92)	0.151 (7.14)	0.134 (1.94)
Nongovernment Sales	0.027 (33.4)	0.034 (20.4)	0.017 (5.73)	0.040 (3.04)

Note: All regressions include year dummies. The within regressions also include firm dummies; the total regressions do not. The weight used is the reciprocal of sales.

competitive procurement is a response to future procurement. The fact that the sum of the  $t$  and  $(t+1)$  coefficients in panel C is close to the corresponding  $t$  coefficient in panel B suggests that the purely contemporaneous version of the model provides reasonable estimates of the long-run responses. Since the coefficient standard errors are much lower in this version, we henceforth exclude leading values of the regressors.

The second extension we consider concerns the distinction between *R&D* and non-*R&D* government contracts. Previous econometric studies of the effect of government procurement on private *R&D* investment have examined the effect only of *R&D* contracting, and not of contracting for non-*R&D* services and products.<sup>14</sup> Moreover, these studies have not distinguished between competitive and noncompetitive procurement of *R&D*. For both reasons, the models upon which these studies were based may have been misspecified and the empirical results, misinterpreted. In panel D we present estimates of the model in which both competitive and noncompetitive contracts are disaggregated into *R&D* and non-*R&D* contracts. Focusing again on the instrumental variable total estimates, we find that both *R&D* and non-*R&D* competitive contracts have a large positive effect on private *R&D*. It is somewhat surprising that the coefficient on the latter is larger (since most of these are price-competitive contracts), but the difference is far from significant:  $F_{1,1003} = 0.14$  ( $p$ -value = .712). The point estimates of the coefficients on both *R&D* and non-*R&D* noncompetitive contracts are negative, but the non-*R&D* coefficient is small and insignificant, whereas the *R&D* coefficient is very large and significantly different from zero. The latter is also significantly different from both the competitive *R&D* coefficient ( $P$ -value = .084) and from the noncompetitive non-*R&D* coefficient ( $P$ -value = .040). To summarize the implications of this model: competitive contracts (whether or not for *R&D*) have a large positive effect on private

*R&D*, noncompetitive *R&D* contracts have an even larger negative ("crowding-out") effect, and other noncompetitive procurement has essentially no effect.

Although the hypotheses that competitive and noncompetitive non-*R&D* have equal effects on private *R&D*, and that competitive and noncompetitive non-*R&D* have equal effects, are rejected by the data, to achieve comparability with previous studies it is useful to impose these restrictions; we do so in panel E. The instrumental variable total estimates suggest that the net effect of government *R&D* contracting on private *R&D* investment is negative: the negative effect of noncompetitive *R&D* contracting outweighs the positive effect of competitive *R&D* contracting. A *ceteris paribus* increase of \$1 in the value of *R&D* contracts would evidently result in a 48-cent reduction in private *R&D* expenditure. In contrast, the positive effect of competitive non-*R&D* procurement outweighs the negative effect of noncompetitive non-*R&D* contracts, so that the net impact of non-*R&D* procurement is positive. Because *R&D* accounts for only about one-sixth of the total value of contracts, government procurement as a whole has a positive and substantial effect on private *R&D* investment.

#### IV. Summary and Conclusions

Competitive procurement, of both *R&D* and other services and products, stimulates considerable private *R&D* investment. Because the firm that is awarded the initial competitive contracts for a weapons system is virtually guaranteed to receive a stream of noncompetitive follow-on contracts, the amount of private *R&D* investment associated with competitive procurement is large relative to the value of competitive contracts. A \$1 increase in competitive procurement is estimated to induce 54 cents of additional private *R&D* investment.

Noncompetitive *R&D* procurement tends to crowd out private *R&D* investment. The award of noncompetitive *R&D* contracts signals the end of the design and technical competition. At this stage of the procurement cycle, there are incentives for firms to

<sup>14</sup> See, for example, David Levy and Nestor Terleckyj, 1983, and John Scott, 1984.

reduce private *R&D*. Losers of the competition reduce spending because the prize is no longer at stake; the winner reduces spending because the government is now willing to directly sponsor the *R&D* via contracting. A \$1 increase in noncompetitive *R&D* procurement tends to reduce private *R&D* by more than \$2. Noncompetitive non-*R&D* has essentially no effect on private *R&D* investment.

In contrast to previous studies of the effect of government *R&D* on private *R&D*, which have not controlled for non-*R&D* procurement and which have not distinguished competitive from noncompetitive procurement, we find that the net effect of *R&D* procurement on private *R&D* is negative. But non-*R&D* procurement, which is about five times as large, has a stimulatory effect, so the net effect of procurement in general is positive and quantitatively important. Over the course of the defense buildup that occurred during the period 1979–84, the increase in government sales accounted for about one-sixth of the total increase in demand. (The government's share in (the level of) sales tends to be lower, about 5 to 8 percent.) Our estimates imply that slightly over half of the total induced increase in private *R&D* was induced by the increase in government demand. The government therefore appears to play a larger role in determining the allocation of the nation's scientific and technical resources, hence the rate and direction of technical progress, than is perhaps generally recognized.

#### DATA APPENDIX

The data used in the econometric analysis were derived from two sources: a computer tape prepared for the author by the Federal Procurement Data Center, and the widely available *Compustat General Annual Industrial File* distributed by Standard & Poor's.

The Federal Procurement Data Center, which is part of the General Services Administration, has since 1978 administered the Federal Procurement Data System, a uniform system for reporting data on procurement by federal agencies.<sup>15</sup> The computer tape provided by the

Federal Procurement Data Center contained a putatively complete enumeration of every contract action during 1979–85 corresponding to the top 1500 (ranked by value of contract actions) contractors; the tape contained records of about 1.3 million contract actions. Each record included the following data on the attributes of the contract action:

- 1) The calendar year in which the action occurred (*YEAR*),
- 2) A numerical code assigned by Dun and Bradstreet to identify the "ultimate parent corporation" of the contractor (*DUNS*)
- 3) the number of dollars obligated or deobligated by the action (*VALUE*)
- 4) The method of procurement associated with the action (*METHOD*)
- 5) A four-character product or service code (*PSC*).

The rules for determining whether a contract action is (a) competitive and (b) for *R&D* were as follows:

if *METHOD* = 3 ("negotiated competitive"), then the action is competitive (*COMPET* = 1); otherwise it is noncompetitive (*COMPET* = 0)

if the first character of *PSC* was an "A," then the contract is for *R&D* (*RD* = 1); otherwise it is not for *R&D* (*RD* = 0).

We computed the total value of each firm's contract actions, by method (competitive versus noncompetitive) and product (*R&D* versus non-*R&D*) in each year by aggregating *VALUE* by *YEAR*, *DUNS*, *COMPET*, and *RD*.

The following were the data items obtained from the *Compustat* file:

- 1) Data (fiscal) year (*YEAR*)
- 2) *CUSIP* Company Number (Issuer Code) (*CUSIP*)
- 3) Sales—Net (*SALES*)
- 4) (Company-funded) *R&D* expense (*CRD*)
- 5) Industry classification number (*DNUM*)

Because the Federal Procurement Data Center and *Compustat* use different schemes for coding companies, a concordance between the *DUNS* and *CUSIP* schemes was required to merge the two files; this was obtained from Harvard Business School. Most of the (smaller) companies included in the Federal Procurement Data Center file are not publicly traded companies and are therefore excluded from the *Compustat* file. Only companies represented in both files (and with nonmissing data on *SALES* and *CRD*) were included in our sample. *NONGOV* was defined as *SALES* minus the total value of contract actions.

The instrument used for the value of competitive contracts awarded to firm *i* in year *t* is the value of competitive contracts that were "potentially awardable" to firm *i* in year *t*. "Potentially awardable" contracts *PCOMP<sub>it</sub>* is defined as the total (across all firms) value in year *t* of competitive contracts for two-digit Federal Supply Code products and services that the firm ever sold to the government during the period 1979–85. This may be expressed algebraically as follows:

$$PCOMP_{it} = \sum_j D_{ij} * AGG_{jt},$$

<sup>15</sup>See the Federal Procurement Data Center publications cited.



where  $D_{ij}=1$  if firm  $i$  ever sold product  $j$  to the government during the period 1979 to 1985;  
 $=0$  otherwise.

$AGG_{jt}$  = total (across firms) value of competitive contracts for product  $j$  in year  $t$ .

(It seems reasonable to maintain that both the lines of business the firm was engaged in, and the aggregate volume of government procurement in those markets, is exogenous with respect to the firm's rate of R&D investment.) An analogously defined instrument is used for noncompetitive contracts. The instrument for  $NON-GOV_{it}$  is the aggregate value of SALES in year  $t$  of all firms in the *Compustat 1987 Annual Industrial File* with the same industry classification number (DNUM) as firm  $i$ .

A complete list of the CUSIP numbers of the firms included in the sample is available from the author and is on file with the *American Economic Review*.

## REFERENCES

- Danhof, Charles H., *Government Contracting and Technical Change*, Washington: The Brookings Institution, 1968.
- Griliches, Zvi, "Productivity, R&D, and Basic Research at the Firm Level in the 1970's," *American Economic Review*, March 1986, 76, 141-54.
- \_\_\_\_\_, and Hausman, Jerry, "Errors in Variables in Panel Data," *Journal of Econometrics*, February 1986, 31, 93-118.
- \_\_\_\_\_, and Lichtenberg, Frank, "R&D and Productivity at the Industry Level: Is There Still a Relationship?," in Z. Griliches, ed., *R&D, Patents, and Productivity*, Chicago: University of Chicago Press, 1984.
- Kihlstrom, Richard E. and Riordan, Michael, "Advertising as a Signal," *Journal of Political Economy*, June 1984, 92, 427-450.
- Levy, David and Terleckyj, Nestor, "Effects of Government R&D on Private R&D and Productivity: A Macroeconomic Analysis," *Bell Journal of Economics*, Autumn 1983, 14, 551-61.
- Lichtenberg, Frank, "The Duration and Intensity of Investment in Independent Research and Development Projects," *Journal of Economic and Social Measurement*, October 1986, 14, 207-18.
- Mansfield, Edwin, *Technological Change*, New York: W. W. Norton, 1971.
- Nalebuff, Barry and Stiglitz, Joseph, "Prizes and Incentives: Towards a General Theory of Compensation and Competition," *Bell Journal of Economics*, Spring 1983, 14, 21-43.
- Reppy, Judith, "Defense Department Payments for 'Company-Financed' R&D," *Research Policy*, 1977, 6, 396-410.
- Roberts, Edward, "How the United States Buys Research," in David Allison, ed., *The R&D Game: Technical Man, Technical Managers, and Research Productivity*, Cambridge: MIT Press, 1969, 280-96.
- Scherer, F. M., "Using Linked Patent and R&D Data to Measure Interindustry Technology Flows," in Z. Griliches, ed., *R&D, Patents, and Productivity*, Chicago: University of Chicago Press, 1984, 417-61.
- Scott, John T., "Firm versus Industry Variability in R&D Intensity," in Z. Griliches, ed., *R&D, Patents, and Productivity*, Chicago: University of Chicago Press, 1984, 233-45.
- Spence, A. Michael, *Market Signaling: Informational Transfer in Hiring and Related Screening Processes*, Cambridge: Harvard University Press, 1974.
- White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, 48, 817-38.
- Department of Defense, *Prime Contract Awards, Fiscal Year 1984*, Washington: USGPO, 1985.
- Federal Procurement Data Center, (1982a), *Reporting Manual*, Arlington, VA, October 1982.
- \_\_\_\_\_, (1982b), *Product and Service Codes*, Arlington, VA, October 1982.
- \_\_\_\_\_, *Standard Report*, Arlington, VA, January 1988.
- National Science Board, *Science Indicators/The 1985 Report*, Washington: USGPO, 1985.
- U.S. Congress, Senate, Committee on the Judiciary, "Competition in Defense Procurement, Hearings before the Subcommittee on Antitrust and Monopoly," Washington: USGPO, 1969.

# Optimal Highway Durability

By KENNETH A. SMALL AND CLIFFORD WINSTON\*

The U.S. highway system represents the nation's largest civilian public investment. Its recent deterioration and rising consumption of public expenditures, now some \$60 billion annually,<sup>1</sup> have generated interest in new tax instruments, pavement management strategies, and rehabilitation methods. Researchers from many disciplines, including economics, have responded with proposals such as marginal-cost user fees and optimal pavement maintenance and repair schedules.<sup>2</sup> A more fundamental issue, highway durability, remains the province of engineers who, through a combination of experimentation, experience, and judgment, have established detailed guidelines for pavement design (American Association of State Highway and Transportation Officials, 1981, 1986).

The purpose of this paper is to apply economic analysis to the durability problem so long dominated by engineering approaches. We formulate a simple cost model in which highway durability is a long-run decision variable. We then apply this model, along with empirical data, to sets of conditions typical of many important U.S. highways, solving in each case for the optimal

level of durability. In the process, we re-analyze the results of a large-scale pavement experiment that has greatly influenced pavement design throughout the world, finding that modern econometric techniques produce some very important differences that may explain the unexpectedly rapid pavement deterioration of U.S. highways.

The results have dramatic implications for highway-building practice and for the role of the trucking industry in highway finance. We find that in order to minimize discounted lifetime costs, typical high-volume urban interstate highway pavements should be more durable and require fewer resurfacings than at present, for a possible net saving of 40 percent of the present value of outlays on resurfacing. Furthermore, although existing highways have marginal pavement-wear costs that are quite high, optimal high-volume urban interstates would not. Thus the need for marginal-cost taxation and the accompanying diversion of trucking-industry revenues would be virtually eliminated on an important segment of the nation's highway network if it were built to optimal standards.

## I. Model

We consider a one-mile, one-directional stretch of highway of width  $W$  (measured in number of lanes). It incurs annual traffic of  $Q_i$  passages by vehicles in type-weight class  $i$ . Each vehicle in class  $i$  contributes as much to road wear as  $l_i$  single axles weighing 18,000 pounds: it is said to have  $l_i$  equivalent single-axle loads (ESALs). We follow common practice in ignoring variation in these equivalence factors with terrain, climate, and highway design. We therefore define annual traffic loadings  $Q = \sum_i l_i Q_i$ , and assume that maintenance cost is an increasing function of  $Q$ .

Highways are built with greater durability in order to reduce the road wear caused by traffic loadings. Durability may be increased

\*Department of Economics, University of California, Irvine, CA 92717, and Brookings Institution, Washington, D.C. 20036, respectively. This work was supported by the Institute of Transportation Studies, University of California, and the Brookings Institution. The results and views expressed are those of the authors and not necessarily of any institution with which we are associated. We are grateful to Robert Crandall, Raymond Forsyth, Ann Friedlaender, Tony Gómez-Ibáñez, David Lilien, David Luhr, Steve Morrison, David Newbery, David Starkie, Peter Swan, and the referees for helpful comments. We also thank Heping He for research assistance.

<sup>1</sup>Motor Vehicle Manufacturers' Association, *Motor Vehicle Facts and Figures*, 1986, p. 83.

<sup>2</sup>For example, D. W. Potter and W. R. Hudson (1981); U.S. Federal Highway Administration (1982, Appendix E); José Gómez-Ibáñez and Mary O'Keeffe (1985); Kenneth Small and Clifford Winston (1986a).

in many ways, including better materials, drainage, and construction techniques; but the most common is thicker pavements or base materials. Hence we define  $D$ , our measure of road durability, to be pavement thickness (or a weighted combination of various component thicknesses) in inches. Greater  $D$  increases capital cost but decreases maintenance cost. We assume that all pavement deterioration is due to traffic loadings.<sup>3</sup>

We therefore write annualized highway maintenance and capital costs as

$$(1) \quad m = rM(Q, W, D),$$

$$(2) \quad k = rK(W, D),$$

where  $r$  is the interest rate,  $M$  is the present discounted value of all required highway maintenance expenses, and  $K$  is the capital cost of construction. Short-run marginal-cost pricing of traffic loadings, at any level of  $D$ , would imply a price equal to

$$(3) \quad \text{SRMC} = \partial(m + k)/\partial Q \\ = r\partial M/\partial Q.$$

Optimal durability is determined by minimizing the sum of  $m$  and  $k$  with respect to  $D$ .

We now specify the maintenance cost in more detail. The dominant component is the cost of periodic resurfacing. As a good approximation, a pavement can be considered to have a lifetime  $N(D)$  giving the number of ESALs that can pass over it before it must be resurfaced (U.S. Federal Highway Ad-

ministration (FHWA), 1982, p. IV-42).<sup>4</sup> Standard practice requires building all lanes of a highway to the same thickness, and repaving all lanes as soon as one of them, normally the outer lane, needs it (Gómez-Ibáñez and O'Keeffe, 1985, p. C-10). The number of traffic loadings between resurfacings is therefore  $N(D)$  divided by the proportion  $\lambda$  of traffic loadings that occur in the outer lane. (We ignore here the option of strengthening a pavement at the time of resurfacing.) Letting  $C(W)$  be the cost of resurfacing,  $M$  is the present value of an infinite sequence of expenditures each of amount  $C(W)$ , made every  $T$  years beginning at time  $T$ ,

$$(4) \quad M(Q, W, D) = C(W)/(e^{rT} - 1),$$

where

$$(5) \quad T = N(D)/(\lambda Q).$$

The wear-related user charge per ESAL mile, from (3)–(5), is then

$$(6) \quad \text{SRMC} = \frac{\alpha C(W)\lambda}{N(D)},$$

where  $\alpha \equiv (rT)^2 e^{rT}/(e^{rT} - 1)^2$ . Hence SRMC is proportional to the undiscounted resurfacing cost per traffic loading, with proportionality constant  $\alpha$  lying between zero and one.<sup>5</sup> Hereafter we refer to (6) as simply the marginal cost.

Following several previous studies, we assume that construction cost is a linear func-

<sup>3</sup>This is a standard assumption: see, for example, U.S. Federal Highway Administration (1982). There is only equivocal evidence that age and weather affect pavements independently from traffic, and that evidence is primarily for thin asphalt pavements; see William Paterson (1987) and the critique in our 1988 manuscript. Including such effects would further strengthen our conclusion that pavements are too thin. This is because age-related deterioration shortens the life of a pavement without affecting the marginal gain from increased durability: that gain is therefore realized sooner and hence discounted less.

<sup>4</sup>A model of optimal maintenance policy would let the level of pavement deterioration that triggers resurfacing be endogenous, influenced by the tradeoff between resurfacing costs and user costs: see Gómez-Ibáñez and O'Keeffe (1985). However, such a fine-tuned maintenance strategy is difficult to achieve in practice, and including it would not alter our results much.

<sup>5</sup>To see this, write  $x = rT$  and  $\alpha = x^2/(e^x + e^{-x} - 2)$ . Use l' Hôpital's rule to find the limit as  $x$  goes to 0, and expand the exponentials in the denominator as power series to see that the denominator is always greater than  $x^2$ .

tion of width.<sup>6</sup> The portion that varies with width includes the cost of grading, which does not depend on pavement thickness, and the cost of the pavement itself, which is approximately proportional to the volume of paving material. Hence a plausible specification is

$$(7) \quad K(W, D) = k_0 + k_1 W + k_2 WD.$$

To find optimal durability  $D^*$ , we numerically minimize  $[M(Q, W, D) + K(W, D)]$  with respect to  $D$ . Dropping terms that are unaffected by  $D$ , and dividing by  $W$ ; this amounts to minimizing total pavement cost (TPC) per lane,

$$(8) \quad \text{TPC} \equiv \frac{C(W)}{W} \cdot \frac{1}{(e^{rT} - 1)} + k_2 D,$$

with  $T$  depending on  $D$  through equation (5).

## II. Empirical Parameters

We consider a six-lane urban interstate highway ( $W = 3$ ), for which the U.S. FHWA (1983, p. II-16) assumes the fraction of trucks in the outer lane to be  $\lambda = 0.7$ . The maintenance cost  $C(W)/W$  is taken to be the per lane cost of repaving a six-lane expressway in an outlying urban residential area, as given in U.S. FHWA (1983, p. II-10) and adjusted to 1984 prices using the FHWA Construction Cost Index (U.S. FHWA, 1985a, Table PT-1). The result is \$113,400 per lane-mile, a conservative figure because it excludes disruption costs.

*Pavement Life and Pavement Thickness:*  $N(D)$ . The relation between pavement life  $N$  and thickness  $D$  was studied as part of a major road test carried out by the American Association of State Highway Officials (AASHTO) between 1958 and 1960. This test remains today the most important and widely used source of experimental in-

formation on the subject. Using test tracks in northern Illinois, investigators measured the effects of axles of various weights on pavement deterioration. The major portion of the experiment consisted of observing pavement quality over a two-year period for each of 548 distinct combinations of axle weight, axle configuration, and road design. Each experimental section received up to 1.1 million passes of the particular axle assigned to it, and its condition was measured periodically. The data are published in Appendix A of Highway Research Board (1962, pp. 244-48, 273-77).

Two kinds of pavements were studied: rigid (portland cement concrete) and flexible (bituminous concrete, commonly called asphalt). The former is more common for heavy-duty roads. The results were incorporated into the standard pavement design guide on which most states base their design practice.<sup>7</sup> The design guide suggests adjustments for different soils and climates; hence our results apply directly only to soils and climates similar to those of the test tracks.

AASHTO specified a nonlinear equation relating a precisely defined measure of pavement quality,  $\pi$  to the number of applications  $n$  of an axle of weight  $L_1$  (in thousands of pounds) and type  $L_2$  ( $L_2 = 1$  for single axles, 2 for tandem). For flexible pavements, where seasonal differences in pavement vulnerability are large,  $n$  is a seasonally weighted number of applications, the weights having been developed from a separate analysis. (This means that our results for the user charge on flexible pavements must be interpreted as pertaining to a seasonally varied charge.) The equation is

$$(9) \quad \pi = \pi_0 - (\pi_0 - \pi_f)(n/\rho)^\beta,$$

where  $\pi_0$  is initial pavement quality;<sup>8</sup>  $\pi_f$  is a predetermined "terminal" pavement quality

<sup>7</sup>See American Association of State Highway and Transportation Officials, 1981, 1986; U.S. FHWA, 1982, p. IV-42.

<sup>8</sup>This is 4.5 for rigid pavements, and 4.2 for flexible, based on average measurements of new pavements.

<sup>6</sup>John Meyer, John Kain, and Martin Wohl, 1965; Marvin Kraus, Herbert Mohring, and Thomas Pinfeld, 1976; Kraus, 1981.

at which the pavement is considered to be worn out;<sup>9</sup> and  $\rho$  and  $\beta$  depend parametrically on  $L_1$ ,  $L_2$ , and  $D$  as described below. By setting  $n = \rho$ , we see that  $\rho$  is just the number of axle passes that will cause the pavement to wear out.

For each of the 548 pavement sections, parameters  $\beta$  and  $\rho$  were estimated by ordinary least squares applied to (9) in the form

$$(9a) \quad \log\{(\pi_0 - \pi)/(\pi_0 - \pi_f)\} \\ = [-\beta \log \rho] + [\beta] \log(n),$$

with the square brackets indicating the intercept and slope of the regression. However, bounds were placed on permissible values of  $\beta$ .

The AASHO researchers then specified  $\beta$  and  $\rho$  parametrically as

$$(10) \quad \beta = b_0 + B_0(D+1)^{-B_1} \\ \times (L_1 + L_2)^{B_2} (L_2)^{-B_3},$$

$$(11) \quad \rho = A_0(D+1)^{A_1} \\ \times (L_1 - L_2)^{-A_2} (L_2)^{A_3},$$

where  $b_0$  was predetermined through a somewhat arbitrary and unclear procedure. For rigid pavements,  $D$  is just the pavement thickness in inches. For flexible pavements,  $D$  is a linear combination of pavement, base, and subbase thicknesses with coefficients .44, .14, and .11; it is known as structural number. The  $A$ 's and  $B$ 's were estimated separately for rigid and flexible pavements on the cross sections of experimental pavement sections (264 rigid, 284 flexible), using the previously estimated values of  $\beta$  and  $\rho$  as dependent variables. In the case of flexible pavements, the coefficients defining the

structural number were estimated along with other parameters of (10) and (11) in a complex multistage procedure (Highway Research Board, 1962, pp. 36-40).

The AASHO estimates of these equations are given in Highway Research Board (1962, pp. 40, 152). Recall that  $N(D)$  is defined in this paper as the number of single axles of weight 18,000 pounds whose passage lowers  $\pi$  to the critical value triggering resurfacing, which we take to be 2.5. Hence to compute  $N(D)$  based on AASHO's equations, we substitute the values  $L_1 = 18$  and  $L_2 = 1$  into (10) and (11), then solve (9) for the value of  $n$  that yields  $\pi = 2.5$ .

*An Alternative Estimate of  $N(D)$ .* Unfortunately, AASHO's functional specification and statistical estimation of the coefficients in equations (9)-(11) were seriously flawed. Problems include mismatched units, an inefficient sequential estimation procedure, extremely poor fits of equation (9) that tended to overstate lifetimes, and arbitrary procedures to compensate for those poor fits; see Small (1985) and Paterson (1987, Sec. 9.3.1) for full critiques. Our interest here is only equation (11). Its left-hand side is observed directly, permitting consistent and efficient reestimation of that equation using Tobit analysis.<sup>10</sup> We use the same functional form as the original AASHO researchers; then we perform a sensitivity analysis using alternative functional forms.

Our estimates of (11) are shown along with AASHO's in Table 1. Recall that the dependent variables are slightly different, so the two sets of estimates are not strictly comparable. Nevertheless, some important differences are apparent. First, the last row shows that our estimation procedure fits the relevant data points better than AASHO's. Second, our estimates show a somewhat less

<sup>9</sup>The AASHO researchers chose  $\pi_f$  to be 1.5, representing a very badly deteriorated pavement; whereas resurfacing is usually recommended at  $\pi = 2.5$  or even higher.

<sup>10</sup>Many pavement sections outlasted the experiment, so our observed dependent variable is  $\max\{\rho, n_{\max}\}$ , where  $n_{\max}$  is the number of axle passes applied during the experiment. We used the censored regression (Tobit) estimation package in SHAZAM, with censoring at an upper limit of  $n_{\max}$ . Statistical efficiency could, in principle, be increased by estimating (9)-(11) as a system; but only at the risk of serious specification error due to problems discussed in Small (1985).

TABLE 1—ESTIMATES OF EQUATION (11)

Variable	Coefficient	Rigid Pavements		Flexible Pavements <sup>a</sup>	
		Ours <sup>b</sup>	AASHO <sup>c</sup>	Ours <sup>b</sup>	AASHO <sup>c</sup>
constant	$\ln A_0$	13.505 (.307)	13.47	12.062 (.237)	13.65
$\ln(D+1)$	$A_1$	5.041 (.329)	7.35	7.761 (.245)	9.36
$-\ln(L_1 + L_2)$	$A_2$	3.241 (.260)	4.62	3.652 (.147)	4.79
$\ln(L_2)$	$A_3$	2.270 (.242)	3.28	3.238 (.189)	4.33
Number of observations		264		284	
No. of censored observations <sup>d</sup>		191		45	
Standard error of regression <sup>e</sup>		.367		.651	
Standard error of prediction <sup>f</sup>		.370	.515	.629	.673

<sup>a</sup>Using seasonally weighted axle applications.

<sup>b</sup>Estimated using the Tobit model of error structure. Dependent variable is the natural logarithm of the number of axle applications to pavement serviceability index of 2.5. Standard errors are given in parentheses.

<sup>c</sup>From Highway Research Board (1962, pp. 40, 152); the intercept has been converted from base 10 to natural logarithms. Dependent variable was the natural logarithm of an estimated parameter representing the number of axle application to pavement serviceability index of 1.5. Standard errors were not reported.

<sup>d</sup>A censored observation is one for which only a lower bound on the dependent variable is observed, due to the finite duration of the test.

<sup>e</sup>Estimated standard error of the error term in the log form of (11), obtained as part of the Tobit maximum-likelihood parameter estimation.

<sup>f</sup>Root mean squared deviation between predicted and observed logarithm of number of axle loads yielding  $\pi_f = 2.5$ , for those observations where  $\pi_f = 2.5$  was reached (73 such observation for rigid pavements, 239 for flexible). For the "AASHO" column, the AASHO estimate of all three equations, (9)–(11), was used in the calculation.

steep relationship between pavement life and axle load—closer to a third-power law than to the fourth-power law conventionally used to approximate the AASHO findings. Our estimates also show a somewhat less steep relationship between pavement life and pavement thickness.

The most important difference, however, is seen by calculating  $N(D)$ , the pavement life in ESALs. This is done for our estimates simply by substituting  $L_1 = 18$  and  $L_2 = 1$  into equation (11). Doing so reveals that our estimates imply far shorter pavement lifetimes for thick pavements than do AASHO's—as much as 65 percent shorter (9.3 vs. 26.6 million ESALs) for the standard 10-inch rigid slab used on most interstates.

This appears to be the first satisfactory explanation for mounting evidence that the AASHO results overstate lifetimes of thick pavements. An early critique noted that AASHO's estimated curves for thick flexible pavements fit the data poorly (Canadian Good Roads Association, 1962, p. 130). Two studies of heavy-duty rigid pavements in Illinois, one of them involving sections of the road test itself which were incorporated into Interstate Route 80, found the AASHO equations overpredicted actual pavement lives by factors of two to three (Robert Elliott, 1981; Bob Welsh et al., 1981); the same was not true of thinner rigid pavements. Hence the widely noted premature deterioration of interstate pavements cannot

be explained by weather or bad traffic forecasts; we offer a sound basis for revising highway-design practice.

It is nevertheless important to recognize that the estimated lifetimes for thick pavements, whether from our procedure or AASHO's, involve extrapolation well beyond the range of direct observation. The road test was run for just over a million axle passes, whereas a thick pavement typically lasts six to ten million passes of a standard 18,000-pound truck axle. In order to assess the robustness of this extrapolation, and more generally to test the sensitivity of our results, we tried a variety of other specifications of equation (11) including redefining the dependent variable by choosing  $\pi_f = 1.5$ ; replacing  $(D + 1)$  by  $D$  and  $(L_1 + L_2)$  by  $L_1$ ; using a translog functional form; and segmenting the sample by pavement thickness. For rigid pavements, the results were not affected much by any of these changes, adding to our confidence in the conclusions we draw. For flexible pavements, results varied considerably so we regard them as preliminary.<sup>11</sup>

**Other Parameters.** The construction cost parameter  $k_2$  is derived from the average contract price for either portland cement concrete (rigid) or bituminous concrete (flexible), delivered and spread in place. Conversations with highway engineers and an asphalt-company official indicated that this is probably an accurate reflection of the marginal cost of added pavement thickness. Cost per unit of delivered material is given in U.S. FHWA (1985b, p. 2); for bituminous concrete, we assume a density of 130 pounds per cubic foot, then divide by 0.44 to obtain cost per square foot per unit increase in  $D$  (as explained earlier,  $D$  is linear in pavement thickness with coefficient 0.44). We assume a standard 12-foot lane width, and add 20 percent for overhead. The result, per

lane-mile per unit of  $D$ , is \$12,800 for rigid pavements and \$24,820 for flexible ones. Other sources were consulted to verify that these costs are reasonable.

The interest rate  $r$  should represent the alternative real cost of public funds, which are at least partly obtained by diverting private-sector funds earning the before-tax corporate real rate of return. Although real rates of return on bonds historically have averaged less than 4 percent, real pretax rates of return in the private sector are much higher. To be conservative, we use a high value, 9 percent, and check for sensitivity within the range 6–12 percent. Using lower interest rates would further strengthen our main conclusion about optimal durability.

### III. Results: Durability

Table 2 shows optimal durability as a function of constant annual traffic loadings  $Q$ . (Since pavement quality depends on cumulative loadings, incorporating traffic growth into our model would affect neither the marginal tradeoffs involved in minimizing per lane pavement cost nor the comparisons between our results and AASHO's.) The three values of  $Q$  shown are roughly the 5th, 50th, and 95th percentile values for six-lane interstate-highway mileage in the United States.<sup>12</sup> Several points deserve attention.

First, our equations imply optimal rigid-pavement thicknesses that are approximately 1 to 3 inches greater than obtained using the AASHO pavement-wear equations in the same optimization model. This is a result of the different predicted pavement lifetimes already discussed. The shorter pavement life  $T$  predicted by our equation at any given pavement thickness  $D$  means that the benefits from a marginal increase in thickness are reaped sooner and are therefore less heavily discounted. (This effect dominates the fact that the percentage increment in pavement life arising from that marginal increase is

<sup>11</sup>We suspect that the *ad hoc* seasonal-weighting scheme used to measure axle passes significantly contributes to the variation in our findings for flexible pavements. No seasonal weights were used with rigid pavements.

<sup>12</sup>These are derived from data in Gómez-Ibáñez and O'Keeffe (1985, pp. 57, C-3, and C-9).

TABLE 2—OPTIMAL DURABILITY

$Q$ (1000s ESALs/yr)	$D^*$ (inches slab thickness or structural number)			
	Rigid Pavements		Flexible Pavements	
	Ours <sup>a</sup>	AASHO <sup>b</sup>	Ours <sup>a</sup>	AASHO <sup>b</sup>
250	8.8	8.0	5.3	5.1
1,000	11.5	9.8	6.4	6.2
2,500	13.8	11.2	7.3	6.9

<sup>a</sup> $N(D)$  is computed using our estimates of modified equation (11), as given in Table 1.

<sup>b</sup> $N(D)$  is computed using AASHO's estimates of equations (9)–(11), as reported in Highway Research Board (1962, pp. 40, 152).

slightly lower using our equations due to the lower exponent on  $D$ .)

Second, our calculations imply that current practice for both rigid and flexible pavements yields substantially less durable roads than would be optimal. The FHWA's Highway Performance Monitoring System, used in annual congressional reports, assumes certain "default" thicknesses for current roads (U.S. FHWA, 1983, p. II-16). The typical "heavy" rigid pavement is 10 inches: about  $1\frac{1}{2}$  inches below optimum for the median traffic level, at which a 10-inch pavement would last only half the optimal 26 years. For flexible pavements, the corresponding definition of "heavy" is  $D = 5.3$ , though other sources suggest a "high-type freeway" pavement with  $D = 5.7$ ; even the latter has less than half the lifetime of the optimal flexible pavement ( $D = 6.4$ ) at the median traffic level.<sup>13</sup>

<sup>13</sup>It is difficult to precisely characterize past or current pavement designs, because there are so many contributing factors and because we lack data on pavement thicknesses classified by traffic levels. Highway engineers tell us that most states use from 9 to 11 inches of rigid pavement for heavily used interstates, though recently some have been adding another inch as a safety factor. The Pennsylvania pavement design procedure calls for 10 inches for heavy rigid pavements, and structural number 5.5 for heavy flexible pavements. FHWA's classification of existing pavements includes a "high-type" flexible pavement for freeways and expressways that appears to have a structural number of about 5.7; although higher structural numbers may sometimes be built, they are rare and some design charts do not

This deficiency in pavement thickness seems mainly attributable, in the case of rigid pavements, to reliance upon the badly estimated AASHO equations: Table 2 shows that if those equations were correct, optimal durability at the median traffic level would be 9.8 inches, very similar to current design. For flexible pavements, however, use of the AASHO equations accounts for only about one-third the difference between optimal and current design; the rest must be due to failure to incorporate economic optimization into design procedures.

A third point is that optimal design is quite sensitive to traffic. Over the tenfold range of traffic loadings shown, optimal rigid-pavement thickness varies from approximately 9 to 14 inches, corresponding to pavement lives of 3.5 to 28 million ESAL applications in the outer lane.

Finally, the results are surprisingly insensitive to the key parameters determining the tradeoff between capital and maintenance costs. At the median traffic level, the interest rate would have to exceed 18 percent, or the incremental pavement cost would have to be three times as large as we estimate, to justify the current 10-inch standard for rigid pavements.

even go as high as 6.0. Hence typical current practice is best characterized as  $D = 10$  for rigid and  $D = 5.7$  for flexible, for high-volume expressways; and lower for lower-volume expressways and other arterials.



TABLE 3—MARGINAL AND TOTAL COSTS OF HIGHWAY WEAR

	Rigid Pavements			Flexible Pavements		
	$D^a$	SRMC <sup>b</sup>	$r \cdot \text{TPC}^c$	$D^a$	SRMC <sup>b</sup>	$r \cdot \text{TPC}^c$
Short Run (at $Q = 1$ million):						
Light	6.5	17.5	61.2	2.0	1275.7	4,251.6
Medium	8.0	6.9	27.9	3.7	39.1	133.7
Heavy	10.0	2.3	15.9	5.3	3.8	20.8
High-type Fwy	10.0	2.3	15.9	5.7	2.2	17.0
Long Run:						
$Q = 250,000$	8.8	2.7	10.9	5.3	2.2	12.3
$Q = 1,000,000$	11.5	0.9	14.4	6.4	0.6	15.1
$Q = 2,500,000$	13.8	0.4	17.3	7.3	0.3	17.1

<sup>a</sup>Slab thickness in inches for rigid; structural number for flexible. Values for short run are from U.S. FHWA (1983, p. II-16), except see text for high-type freeway. Values for long run are from Table 2.

<sup>b</sup>Marginal pavement-wear cost, from equation (6), in cents per ESAL-mile.

<sup>c</sup>Total pavement cost, from equation (8), multiplied by  $r$  and expressed in thousands of 1984 dollars per lane-mile per year.

#### IV. Results: Marginal and Total Cost

Table 3 presents the marginal cost of highway wear from equation (6), at the FHWA default values of  $D$  for light, medium, and heavy pavements;<sup>14</sup> at our best characterization of typical practice for high-volume freeways (see fn. 13); and at our estimated  $D^*$  for the three traffic levels. It also presents annualized total life-cycle pavement costs from equation (8). Again, several comments are in order.

First, the annualized cost saving for a "high-type freeway" pavement at the median traffic level is \$1500–\$1900 per lane-mile, roughly 40 percent of the annualized discounted value of maintenance cost as currently built. Data from U.S. FHWA (1985a, pp. 146, 148, 168) indicate about 62,000 lane-miles of freeways and expressways of that width or more, so a crude estimate of the annual total cost saving from optimizing durability on them is \$105 million. To properly extend this calculation to all U.S. highways is beyond the scope of this paper; but a very conservative estimate is \$1.0 billion,

obtained by assuming the same cost saving<sup>15</sup> for each of the 550,000 lane-miles of urban arterials and collectors. Calculations using more refined data and less conservative assumptions suggest savings of several times this amount (Small, Winston, and Carol Evans, 1988).

Second, the marginal-cost user charge for high-type freeway pavements is 2.2 to 2.3 cents per ESAL-mile: an amount that, for the typical fully loaded five-axle tractor-trailer combination, would extract payments

<sup>14</sup>Short-run marginal cost depends on  $Q$  through  $\alpha$  and equation (5), but not very strongly; we present results for the median  $Q$ .

<sup>15</sup>There is reason to believe that the possible cost savings are in fact much higher. Design guides generally recommend that engineers aim for a given pavement life—20 years for rigid and 10 for flexible—are fairly common in the United States. Our model applied to six-lane freeways at median traffic levels yields lives for the "high-type freeway" designs of about 13.5 years; and optimal lives of about twice that, for both rigid and flexible. We believe that rigid pavements are typically lasting 13.5 rather than the planned 20 years due to errors in the design equations used, and that optimal design would indeed double their lives. Flexible pavements, in contrast, may undergo substantial age-related deterioration; indeed, our equations can be made to predict a 10-year life for  $D = 5.7$  by simply adding an exponential deterioration of about 4 percent per year, a value right in the middle of the range suggested by Paterson (1987, ch. 8). When we do so, optimal life is found to be 20 years, again about twice the actual and yielding an annual cost saving of \$3,200 per lane-mile. This alone could nearly double the estimate in the text.

comparable to current fuel taxes. Marginal costs are much higher on thin pavements, and user charges as high as \$10 per ESAL mile can be justified in extreme cases.

Third, U.S. trucking-industry representatives are correct in claiming that trucks would not be very damaging if pavements were designed optimally in the first place—at least for high-volume roads. At median traffic levels for six-lane urban interstates, the long-run marginal cost is less than one cent per ESAL-mile. But highway pavements are subject to strong durability economies: long-run average and marginal costs decline markedly with traffic loadings. Hence even in a world of optimal capital stock, marginal-cost user charges for highway wear would show great variation among roads with different amounts of traffic.

Because of these durability economies, efficient wear-related user charges would not fully cover the long-run costs of pavement construction and maintenance. Thus there may still be a role for other charges, such as license and registration fees, even in an optimally built and optimally priced road system. A complete analysis of highway finance must consider a transition period from the present to an optimal system, the practical implementation of wear-related user charges, and their relation to congestion charges. These issues are discussed in Small, Winston, and Evans (1986b, 1988) and Newbery (1987a, b).

### V. Conclusion

Using models of pavement deterioration very close to those developed for the AASHO road test, and estimating them from the road-test data, we find evidence that existing design equations overestimate the life of thick pavements. Furthermore, current and past pavement design practice has led to underinvestment in pavement durability due to reliance upon statistically flawed design equations and failure to incorporate economic optimization.

We do not claim that our models capture all the effects that pavement engineers should take into account. Better materials, drainage, and construction practices are all potential

substitutes for thicker pavements. Furthermore, engineers are moving toward thicker pavements because of disappointing experience with past designs. Nevertheless, design manuals still use the incorrect AASHO road-test equations as a starting point, and are not based upon any explicit economic optimization framework. Hence we believe that the models presented here would substantially improve pavement design.

Our results support the growing consensus that heavy vehicles impose very high marginal pavement-wear costs on many existing roads. However, highway pavements also show substantial long-run economies of scale. If marginal-cost pricing were accompanied by optimal investment, the nation's highest-volume roads would have user charges that were lower, for most vehicles, than existing fuel taxes.

### REFERENCES

- American Association of State Highway and Transportation Officials, *AASHTO Interim Guide for Design of Pavement Structures*, 1972, Chapter III Revised, 1981, Washington: AASHTO, 1981.
- , *AASHTO Guide for Design of Pavement Structures*, Washington: AASHTO, 1986.
- Canadian Good Roads Association, *Report of the Observer Committee of the Canadian Good Roads Association on the AASHO Road Test*, 1962.
- Croney, David, *The Design and Performance of Road Pavements*, U.K. Transport and Road Research Laboratory, London: Her Majesty's Stationery Office, 1977.
- Elliott, Robert P., "Rehabilitated AASH(T)O Road Test," *Paving Forum*, Summer 1981, 3-9.
- Gómez-Ibáñez, José A. and O'Keeffe, Mary M., *The Benefits from Improved Investment Rules: A Case Study of the Interstate System*, Research Report R85-2, John F. Kennedy School of Government, Harvard University, Cambridge, 1985.
- Highway Research Board, *The AASHO Road Test, Report 5: Pavement Research*, Special Report No. 61E, 1962.

- Kraus, Marvin, "Indivisibilities, Economies of Scale, and Optimal Subsidy Policy for Freeways," *Land Economics*, February 1981, 57, 115-21.
- \_\_\_\_\_, Mohring, Herbert and Pinfeld, Thomas, "The Welfare Costs of Nonoptimal Pricing and Investment Policies for Freeway Transportation," *American Economic Review*, September 1976, 66, 532-47.
- Meyer, John R., Kain, John F. and Wohl, Martin, *The Urban Transportation Problem*, Cambridge: Harvard University Press, 1965.
- Newbery, David M., (1987a), "Road Damage Externalities and Road User Charges," *Econometrica*, 56, forthcoming 1988.
- \_\_\_\_\_, (1987b), "Road User Charges in Britain," Discussion Paper No. 174, Centre for Economic Policy Research, London, 1987.
- Paterson, William D. O., *The Highway Design and Maintenance Standards Study, Vol. III: Prediction of Road Deterioration and Maintenance Effects: Theory and Quantification*, draft manuscript, The World Bank, Washington, 1987.
- Potter, D. W. and Hudson, W. R., "Optimization of Highway Maintenance Using the Highway Design Model," *Australian Road Research*, March 1981, 11, 3-16.
- Shook, James F. and Finn, Fred N., "Thickness Design Relationships for Asphalt Pavements," in *Proceedings of the International Conference on the Structural Design of Asphalt Pavements*, Conference Executive Committee, eds. and publishers, Department of Civil Engineering, University of Michigan, 1963, 52-83.
- Small, Kenneth A., "Statistical Problems with the AASHO Road Test Analysis," unpublished paper, University of California-Irvine, September 1985.
- \_\_\_\_\_, and Winston, Clifford, (1986a) "Welfare Effects of Marginal Cost Taxation of Motor Freight Transportation: A Study of Infrastructure Pricing," in Harvey S. Rosen, ed., *Studies in State and Local Public Finance*, Chicago: University of Chicago Press for NBER, 1986, 113-28.
- \_\_\_\_\_, and \_\_\_\_\_, (1986b) "Efficient Pricing and Investment Solutions to Highway Infrastructure Needs," *American Economic Review Proceedings*, May 1986, 76, 165-69.
- \_\_\_\_\_, \_\_\_\_\_, and Evars, Carol, *Road Work: A New Highway Policy*, draft manuscript, The Brookings Institution, 1988.
- U.S. Federal Highway Administration (FHWA), *Final Report on the Federal Highway Cost Allocation Study*, Washington: USGPO, May 1982.
- \_\_\_\_\_, *Highway Performance Monitoring System Analytical Process, Vol. II: Technical Manual*, Washington: U.S. FHWA, March 1983.
- \_\_\_\_\_, (1985a), *Highway Statistics 1985*, Washington: USGPO, 1985.
- \_\_\_\_\_, (1985b), *Price Trends for Federal-Aid Highway Construction*, Washington: U.S. FHWA, Second Quarter, 1985.
- Welsh, Bob H., Witzak, Matthew W., Zimmer, Donald C. and Hacker, Daniel G., "Pavement Management Study: Illinois Tollway Pavement Overlays," *Transportation Research Record*, 1981, 314, 34-40.

# Income Redistribution in a Federal System

By WILLIAM R. JOHNSON\*

The question of the appropriate level of government to undertake income redistribution is receiving renewed attention from economists.<sup>1</sup> The standard argument about the issue might be summarized as follows:

1. Interstate migration of taxpayers and transfer recipients raises the cost to voters of income redistribution at the state level above the cost at the federal level.

2. Statewide redistribution confers spillover benefits to residents of other states.

3. Therefore, as compared with a national system, decentralized responsibility for income redistribution yields both less redistribution and less than the socially optimal amount of income redistribution.

The purpose of this paper is only to question step 1 of the above argument; of course, readers who accept the assumptions of the standard argument may conclude that the possible weakening of step 1 of that argument reduces the likelihood that its conclusion holds. The paper argues that the cost of redistribution faced by state residents may, in fact, be lower at the state level than at the national level because their federal tax obligations fall as the state redistributes more income. Thus, part of the cost of redistribution at the state level can be exported to residents of other states. The state's federal tax bill falls when taxable incomes in the state are reduced by increased redistribution; deductibility of state taxes enhances this result but is not necessary for it.

To demonstrate the proposition without adding unnecessary complication, a simple

model of income redistribution is described in which governments at all levels use linear tax-transfer schedules to redistribute income. The single-government case is considered first, then two levels of government are introduced, and finally migration is allowed. It is shown that, without migration, state costs are never greater than national costs of redistribution as long as redistribution reduces total money income. When migration is allowed, the result depends on the elasticity of migration responses compared with the elasticity of labor supply responses.

## I. A Simple Model of Income Redistribution

A state consists of  $n$  persons, indexed by  $i$ , who differ only in their wage rates,  $w_i$ . The amount of labor supplied by each person,  $L_i$ , depends on the parameters of his budget constraint. All income is labor income. Governments redistribute income by taxing labor income with proportional taxes and giving equal per capita payments, demogrant, to each person.<sup>2</sup> After-tax income for each person is given by  $(1-t) \cdot w_i L_i + B$ , where  $t$  is the tax rate and  $B$  is the demogrant payment. Labor supply is a function of the person's net of tax wage rate and the vertical intercept of his budget constraint:<sup>3</sup>

$$(1) \quad L_i = f[(1-t)w_i, B]; \quad \text{where } f_2 \leq 0.$$

Consider the effect on labor supply of a small change in the parameters of govern-

\*Department of Economics, University of Virginia, Charlottesville, VA 22901. Research support from NBER's project on state and local finance is gratefully acknowledged. Helpful comments on earlier versions were received from Charles Brown, Don Fullerton, Robert Moffitt, Edgar Olsen, Harvey Rosen, Jon Skinner, and anonymous referees.

<sup>1</sup>For example, see Charles Brown and Wallace Oates, 1987, and Edward Gramlich, 1985.

<sup>2</sup>This amounts to a simple counterclockwise rotation of the budget constraint around some break-even point. Net taxes are negative for those below the break-even point and positive for those above it. The plan can also be considered a negative income tax financed by positive taxes at the same marginal tax rate.

<sup>3</sup>Nothing in the theory depends on the labor supply function's being the same for every person, but to avoid writing multiple subscripts, the function  $f$  is written as though it applied to everyone.

ment policy,  $t$  and  $B$ ,

$$(2) \quad dL_i = -f_1 \cdot w_i \cdot dt + f_2 \cdot dB.$$

Imposing the reasonable restriction that any change in policy is self-financing, we have the marginal budget constraint

$$(3) \quad n \cdot dB = \sum [w_i L_i dt + t w_i dL_i],$$

where the summation in (3) and all subsequent equations is taken over all persons in the state. Equation (3) simply says that the added demogrants ( $n \cdot dB$ ) must be financed by additional tax revenue, which is the sum of the added tax rate,  $dt$ , applied to current labor income plus the effect of changes in labor supply on tax revenues. Second-order terms are omitted.

We can derive the government's policy constraint by substituting each person's labor supply response, given by (2), into the marginal budget constraint, (3), yielding

$$(4) \quad \frac{dB}{dt} = \frac{\sum [w_i L_i - t \cdot f_1 \cdot w_i^2]}{n - t \sum f_2 \cdot w_i}.$$

Equation (4) gives the size of the incremental demogrant that can be financed by an increase in taxes; this number depends on the existing tax structure, the distribution of wages, and the labor supply functions of everyone in the political jurisdiction. For example, note that when labor supply is totally unresponsive ( $f_1 = f_2 = 0$ ),  $dB/dt$  is just average labor income, which would be the per capita revenue produced by a 100 percent tax. When labor supply does respond to economic incentives,  $dB/dt$  can be either greater than or less than average labor income, depending on whether redistribution raises or reduces total money income. I consider the case in which redistribution reduces total money income on the margin to be the usual case, although theory does not rule out the other possibility.<sup>4</sup>

Finally, it is easy to show that an individual always prefers a higher incremental demogrant ( $dB$ ), holding constant the associated rise in taxes ( $dt$ ) as well as the starting values of tax rates,  $t$ , and his gross labor income,  $w_i L_i$ . In comparing two situations with the same increase in taxes, the case in which  $dB$  is greater implies that the budget constraint for each person is everywhere outside the budget constraint for the case with the smaller incremental demogrant. Hence, when forced to increase redistributive taxes by given amount  $dt$ , every person in the community would prefer to do it in the way which yields the highest per capita additional demogrant,  $dB$ .

## II. Income Redistribution in a Federal System

Now assume that the state described above is one of  $N$  identical states in a federation. Both the state and federal governments impose proportional taxes to finance demogrants. The tax rate imposed on a resident of the state is the sum of the federal rate,  $t_f$ , and the state rate,  $t_s$ ; the demogrant received is  $B = B_s + B_f$ . Because each state is identical, initial levels of state taxes and benefits are equal in all states. It is assumed, for now, that no one can migrate between states.

Consider the implication to the residents of a particular state of increasing redistribution at the state level. The marginal budget constraint facing the state still restricts incremental demogrants to equal additional state tax revenue

$$(5) \quad n \cdot dB_s = \left( \sum w_i L_i \right) dt_s + t_s \cdot \left( \sum w_i \cdot dL_i \right).$$

Increased redistributive activity by a state will also affect federal revenues by changing the tax base in the state; the federal government is assumed to respond by reducing its

<sup>4</sup>A necessary but far from sufficient condition for redistribution to raise total income at the margin is that some person's labor supply curve be backward-bending.

If income effects are strong enough among taxpayers, increased taxes will raise labor supply enough to offset the inevitable decline among recipients of net transfers.

demogrant accordingly:<sup>5</sup>

$$(6) \quad n \cdot N \cdot dB_f = t_f \cdot \left( \sum w_i \cdot dL_i \right).$$

For the time being, it is assumed that other state governments do not respond to this state's change in state taxes. Combining (5), (6), and (2), we can write the incremental demogrant ( $dB_s + dB_f$ ) caused by an increase in state taxes,  $dt_s$ :

$$(7) \quad \frac{dB}{dt_s} = \frac{\sum w_i L_i - (t_s + t_f/N) \cdot \sum f_1 w_i^2}{n - (t_s + t_f/N) \cdot \sum f_2 w_i}.$$

Now consider the effect of an increase in the federal tax rate on the demogrant received by the residents of each state. The marginal budget constraint for the federal government is

$$(8) \quad N \cdot n \cdot dB_f = N \cdot dt_f \left( \sum w_i L_i \right) + N \cdot t_f \cdot \left( \sum w_i dL_i \right),$$

while the corresponding constraint for each state government is

$$(9) \quad n \cdot dB_s = t_s \cdot \left( \sum w_i dL_i \right).$$

State governments have to adjust their demogrant because the changing federal tax rate affects taxable income in their jurisdictions. Combining (8), (9), and (2), and recalling that  $dB = dB_f + dB_s$ ,  $dt = dt_s + dt_f$ , we have the demogrant that can be financed by an increase in federal taxes

$$(10) \quad \frac{dB}{dt_f} = \frac{\sum w_i L_i - t \cdot \sum f_1 w_i^2}{n - t \cdot \sum w_i f_2}.$$

Section I derived the result that every person in the state will prefer to redistribute using state taxes rather than federal taxes if

$dB/dt_f$  is less than  $dB/dt_s$ . Using (7), (10), and some tedious algebraic manipulation, it can be shown that a necessary and sufficient condition for this to be true is that

$$(11) \quad n \cdot \sum f_1 w_i^2 - \left( \sum w_i L_i \right) \cdot \left( \sum f_2 w_i \right) > 0.$$

Equation (4) reveals that condition (11) holds if and only if redistribution reduces money incomes; that is, if and only if  $dB/dt < (\sum w_i L_i)/n$ . Hence, we have the basic result that with no migration, no response by other states, and no benefit spillovers, and where redistribution reduces money income, every resident of a state will prefer to redistribute using state taxes as opposed to federal taxes. The reason is, of course, that the reductions in state money income reduce that state's federal tax bill and so some of the cost is exported to other states.<sup>6</sup>

*Responses by Other States.* What happens if the residents of this state think that other states will respond to a change in state taxes with tax changes of their own? For example, suppose that residents believe that an increase in  $t_s$  will be matched by the same increase<sup>7</sup> in  $m$  other states, where  $0 \leq m \leq N-1$ . Then, federal tax revenues will change because money incomes are changing in  $m+1$  states, and equation (7) will be altered. Specifically, the expression  $(t_s + t_f/N)$  in (7) is replaced by  $[t_s + t_f(m+1)/N]$ . The case where  $m=0$  and no other state responds has already been considered. The other extreme is where  $m=N-1$  and all other states respond; in that case, there is no difference between state and federal taxes and so no preference in either direction. For intermediate cases, where some states respond ( $m < N-1$ ), the preference for state redistribution still holds but with weakened force. The smaller the number of states which respond, the greater the preference for state redistribution.

<sup>6</sup>Naturally, if redistribution raised money income, then there would be a preference for federal redistribution.

<sup>7</sup>It does not seem reasonable to assume that one state's increase would be more than matched by changes in other states.

<sup>5</sup>Equation (6) ignores the effect of the changed federal demogrant on labor supply behavior in other states. Incorporating such effects would make the analysis much more complex without changing the ultimate outcome.

*Migration.* If the migration of taxpayers and transfer recipients is allowed, the proposition is weakened and quite possibly reversed. From the point of view of a non-migrating state resident, an increase in state taxes induces some high-income persons to leave and entices some low-income persons to enter the state, both reducing the additional demogrant that can be financed by a given increase in state taxes. The migratory response to federal taxation is presumably much weaker since labor is less mobile internationally than between states. Whether state or federal redistribution is preferred depends on the relative strengths of the migration and tax-exporting effects and can only be answered empirically.<sup>8</sup>

Finally, it is interesting to consider the case in which other states can respond when migration is allowed. The response of other states will presumably weaken the migration effect as it weakened the tax-exporting effect. Migrants respond to differences in taxes and transfers between jurisdictions and if other states respond, these differences are muted. In fact, one can speculate that partial response by other states will weaken the migration effect more than the tax-exporting effect because the states who are likely to respond are the nearby states who exert a greater effect on migration than on federal revenues.

To illustrate, suppose New York increases its tax rate. With no response by other states, high-income people move to New Jersey and Connecticut and low-income people move from those states to New York. But the states most likely to respond are New Jersey and Connecticut because New York's action has reduced their cost of redistributing. The

migratory flows might be greatly reduced by New Jersey and Connecticut's tax increase. However, the tax-exporting effect remains strong because tax increases by New Jersey and Connecticut have a relatively small effect on federal tax revenues.

### III. Conclusion

The tax-exporting effect highlighted here arises when redistribution reduces money income and hence federal tax obligations. If the deductibility of state taxes from the federal income tax were considered, it would strengthen the proposition. The possible response of other states to one state's tax change weakens both the tax-exporting and the migration effects and may weaken the migration effect more.

State redistribution clearly generates positive externalities—benefit spillovers and migration responses. However, the current U.S. fiscal system subsidizes state redistribution through tax exporting, tax deductibility, and grants-in-aid.<sup>9</sup> Whether the subsidy is enough to offset the externality is an empirical question not treated here.

<sup>9</sup>Robert Moffitt (1984) reports that grants-in-aid exert a significant effect on state redistribution expenditures.

### REFERENCES

- Brown, Charles and Oates, Wallace, "Assistance to the Poor in a Federal System," *Journal of Public Economics*, April 1987, 32, 307–30.
- Gramlich, Edward M., "Reforming U.S. Fiscal Arrangements," in D. Rubinfeld and J. Quigley, eds., *American Domestic Priorities*, Berkeley: University of California Press, 1985.
- Johnson, William R., "Marginal Costs of Income Redistribution at the State Level," NBER Working Paper No. 1937, 1986.
- Moffitt, Robert, "The Effect of Grants-in-Aid on State and Local Expenditures: The Case of AFDC," *Journal of Public Economics*, April 1984, 23, 279–306.

<sup>8</sup>In my working paper (William Johnson, 1986), I report the results of some simple simulations which attempted to compute the relative strengths of these two effects. The tax-exporting effect depends both on labor supply responses and the level of marginal tax rates. In those crude simulations, I showed that in many plausible cases the tax-exporting effect outweighs the migration response, but the reader should be cautioned that the simulations rested on many simplifying assumptions such as constant elasticity labor supply functions and proportional state tax systems.

# The Strategic Value of Flexibility: Reducing the Ability to Compromise

By NALIN KULATILAKA AND STEPHEN GARY MARKS\*

Production flexibility is achieved by the ability to change a process from one mode of operation to another. For example, a power generation plant can build in the ability to shift between the use of coal and oil. Manufacturing plants can have the ability to switch between using purchased electricity, co-generated electricity, and natural gas. Tire manufacturers can design processes capable of using either natural or synthetic rubbers. In a wide range of manufacturing applications there exists the ability to shift between labor-intensive and capital-intensive technologies.

It is well known that one of the most significant advantages of flexibility is to provide the production process with an ability to modify itself in the face of uncertainty. Recent studies have addressed the issue of the economic value of flexibility in terms of its *option* value. (See Robert McDonald and Daniel Siegel, 1985, 1986; Scott Mason and Robert Merton, 1985.) Nalin Kulatilaka, 1987, develops a general model of flexibility which synthesizes the above options literature. In these studies, random price realizations trigger the exercise of the option to switch from one mode to another. The option has value only if price is uncertain.

In this paper we examine the *strategic* value of flexibility in a world with incomplete contracting. The study of incomplete contracting is motivated by four factors: 1) In a world of complete contingent contracts, most (but not all) interesting strategic considerations disappear. 2) There is a significant literature that addresses the implica-

tions of bargaining under the assumption of incomplete contracting. 3) There are sound theoretical reasons why we would expect contracts to be incomplete. 4) Empirically, we observe that most contracts are indeed incomplete.

One such market in which incomplete contracting is the norm is the labor market. Typical labor contracts specify wage and working conditions, but allow employment levels and levels and types of capital expenditures to be set by the firm. (There are, of course, rare exceptions.) Other instances in which price and other characteristics are negotiated but quantity is left to the option of the purchaser can be found in the purchases of integrated circuits, natural gas, and automobile parts. For example, it is common for computer manufacturers and integrated-circuit suppliers to fix by contract the engineering specifications and the unit price. Thereafter, the computer manufacturer can choose the quantity of chips to be delivered.

The reason usually cited for the absence of complete contingent contracts is the transaction costs of contracting. For example, first-best contracting in the absence of transaction costs might result in contracts that specify price-quantity-investment schedules for a continuum of demand conditions, a continuum of cost conditions, and that are contingent on the random arrival of new technologies. Cost of negotiation, as well as imperfect or asymmetric information, make such contracts impossible in practice. In addition, complete *noncontingent* contracts, such as one that would guarantee a quantity level over all states of nature, may not represent the second-best solution. Firms facing uncertainty derive an option value from the ability to vary quantity with demand conditions, cost conditions, and the arrival of new technologies. That is, the value of the option

\*Boston University, Boston, MA 02215. Kulatilaka: Department of Finance and Economics, School of Management. Marks: Department of Finance and Economics, School of Management and the School of Law. We wish to thank Carliss Baldwin, Alex Kane, David Lax, William Samuelson, and an anonymous referee for their comments and suggestions.



to the purchaser may be greater than the value of a fixed supply contract to the supplier. In such a case the contract will allow the purchaser to choose quantity.

In addition, information asymmetries may make it impossible to fix quantity. In the case of labor, guaranteed employment (fixed quantity) in the presence of informational asymmetries may lead to problems of moral hazard in the choice of effort by workers or of adverse selection in the retention of high productivity workers.

It is not surprising, then, that incomplete contracting is the rule rather than the exception in many industries. It is not the purpose of this paper, however, to model the causes of incomplete contracting. Rather, we follow Paul Groot (1984) and others and examine the implications of bargaining under the assumption that contracts are incomplete. The advantage of assuming incompleteness is that it allows us to model bargaining under certainty. This not only makes for a more tractable model but it also enables us to isolate the purely strategic elements of a bargain. For example, Groot assumes that wage is bargained (contracted) but that the firm chooses both employment and capital expenditures. Our model is in the same spirit. In order to focus on the strategic aspects of flexibility, we consider a model under certainty where 1) the firm chooses a technology, 2) the firm and its input suppliers (and/or output purchasers) negotiate over price, and 3) the firm subsequently chooses input quantities. (The model developed herein builds on one developed by Stephen Marks, 1984, who studied the strategic implications of wage bargaining.)

The following simple thought experiment illustrates the basic approach. Consider two scenarios. In scenario I there is a fixed technology in place that uses one unit of labor. The firm and the labor union bargain over the pre-wage surplus,  $R$ . In scenario II the firm has a flexible technology in place with two modes. The first mode is identical to the fixed technology of scenario I. The second mode uses less labor. Suppose that it is profitable to switch to the second mode if wage is high. That is, the firm now has a threat that says to labor, "if you demand too

high wages, we will employ fewer workers." Let us suppose that this flexibility is costless to install and that switching between modes is also costless. Finally, suppose that there is no uncertainty regarding the production process, future prices, or bargaining outcomes.

It is well established that such flexibility can confer a strategic advantage on the firm resulting in a lower input price. It is also well known that such flexibility has the potential for inefficiency. That is, it may lower joint profits. For example, Carliss Baldwin (1983) develops a model where building one efficient plant and one inefficient plant (rather than two efficient plants) produces a strategic advantage in wage bargaining since if wage is too high the firm will shut down the inefficient plant. The resultant restraint on wage bargaining allows the firm to capture profits in the efficient plant. The firm is better off but the result is socially inefficient. Another example is the case of an upstream monopolist forcing flexible downstream producers to shift to inefficient technologies. Flexibility makes the downstream firms better off but the result is socially inefficient.

A more striking result is that such flexibility may be detrimental to the firm itself. We develop a model that demonstrates how this may happen. We show that in some cases a mutually destructive (but incentive-compatible) threat may be exercised if in place. In these cases a rational firm will avoid flexibility even if it is technologically costless. Whether flexibility will be detrimental or beneficial (or whether it will have no value) depends on the exact parameters of the problem. We will show, however, that very similar parameters lead to very different results as to the value of flexibility.

## I. The Model

Here we present the outline of the model. The formal propositions and proofs are given in Appendix I.

Consider a fixed technology which uses  $\alpha$  units of input and generates revenues  $A$ . The price of the input,  $P$ , is negotiated after the technology is put into place. The value of this technology to the firm,  $V_A$ , is simply

$(A - \alpha P)$  if operated and zero if closed. (We ignore installation and fixed costs for now.) Another fixed technology uses  $\beta$  units of the input and produces a value,  $V_B$ , equal to  $B - \beta P$  if operated and zero if closed. We can assume, without loss of generality, that  $A \geq B$ . The revenues not received by the firm are received by the input supplier. The input supplier is assumed to maximize net profits

$$(1) \quad W = (P - C)Q,$$

where  $C$  is a unit opportunity cost and  $Q$  is quantity. We will assume for the exposition that  $C = 0$  although this is unimportant.  $Q$  is determined by the firm after bargaining and can be zero,  $\alpha$ , or  $\beta$  depending on the technology in place.

Now consider a flexible technology  $F$  which has as its two modes the fixed technologies  $A$  and  $B$ . The mode of operation of the flexible technology is chosen by the firm after  $P$  has been determined. Hence, its value,  $V_F$ , is given by  $\max(A - \alpha P, B - \beta P)$  if operated and zero if closed. The flexible manufacturing technology will operate in mode  $A$  only if the profits in mode  $A$  are greater than the profits in mode  $B$ , that is, only if

$$A - \alpha P \geq B - \beta P$$

or, equivalently, if

$$(2) \quad P \leq (A - B)/(\alpha - \beta) \equiv P_s,$$

where  $P_s$  is called the switching price for the flexible technology. Otherwise, it will operate in mode  $B$ . (Note that if there is equality we have assumed the firm will operate in mode  $A$ . Also, the firm may choose not to operate if profits are negative.)

A necessary condition for a positive switching price (given  $A > B$ ) is that  $\alpha > \beta$ . (Note that if  $\alpha \leq \beta$ , the firm will choose mode  $A$  irrespective of price. In such a case, the flexible technology is not really flexible. That is, the alternative mode,  $B$ , is not a credible threat as it requires more of the input. In such cases, "flexibility" will have no value. We henceforth consider only cases where  $\alpha > \beta$ . However, as we will see, flexi-

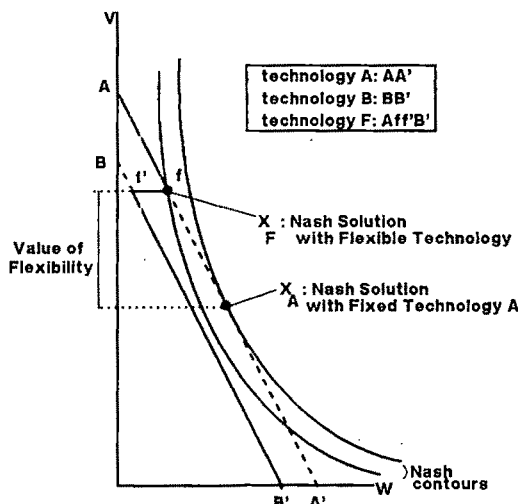


FIGURE 1. FLEXIBILITY IS BENEFICIAL TO THE FIRM

bility may have no value even if  $\alpha > \beta$ .) In Figure 1 we plot the firm's profit,  $V$ , against the input-supplier's revenues,  $W$ , for the three technologies.

We now apply the Nash-bargaining solution to the set of feasible outcomes. The Nash solution is well known (see John Nash, 1950; Alvin Roth, 1979) and its use in labor-firm bargaining is not without precedent (see Grout, 1984). The Nash solution satisfies Pareto optimality, invariance with respect to linear utility transformations, independence from irrelevant alternatives and symmetry. Although the Nash solution is usually used with convex feasible sets, it is Pareto optimal even when the feasible set is not convex. Symmetry implies equal bargaining power. (We relax this assumption in Appendix I.) When applied to a constant surplus, the Nash solution splits the surplus. In our notation the Nash-bargaining solution is to choose a  $P$  that maximizes the product of  $V(P)$  and  $W(P)$ . Nash contour lines are given by  $V(P)W(P) = k$  for various values of  $k$ . In Figure 1, the Nash solution is the intersection of the highest Nash contour line and the feasible set. The Nash solution is  $X_A$  under the fixed technology  $A$  and  $X_F$  under the flexible technology. Clearly, the firm is better off under the flexible technology, that is, flexibility has value.

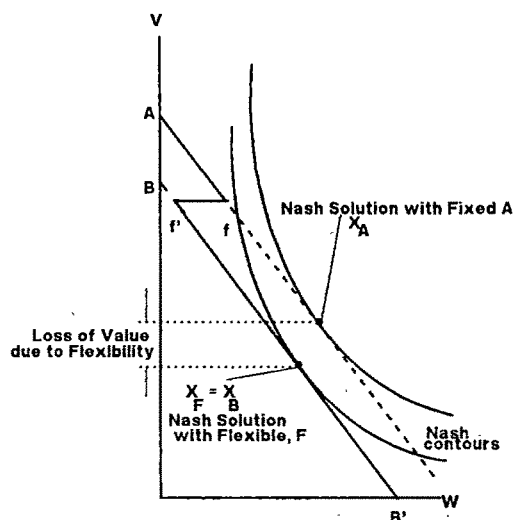


FIGURE 2. FLEXIBILITY IS DETRIMENTAL TO THE FIRM

Figure 2 depicts a situation with slightly different parameter values but a dramatically different result. The flexible technology leads to lower value to the firm. A rational firm will adopt a flexible technology under the parameter values of Figure 1 but a fixed (type A) technology under the parameters of Figure 2. That is flexibility can be detrimental to the firm even if it is costless. The formal propositions are presented in Appendix I.

Heuristically, the reason for these results is that the bargaining is over price but not quantity and flexibility removes from consideration a set of prices. In addition it alters the payoffs for another set of outcomes.

To see this recall that  $P_s$  is the price above which the firm will switch from mode A to mode B if the flexible technology is in place. Let  $P' = P_s \alpha / \beta$ . Now consider the interval  $(P_s, P')$ . Under the flexible technology any price in this interval is Pareto dominated by  $P_s$ . For the firm any price above  $P_s$  results in a worse outcome than  $P_s$ . For the input supplier, a price above  $P_s$  results in a greater unit return but a lower quantity. For prices below  $P'$ , total revenues to the input supplier are lower than with  $P_s$ , that is,

$$\begin{aligned} \text{Revenue at Price } P \text{ in } (P_s, P') \\ = P\beta < P'\beta = P_s\alpha = \text{Revenue at } P_s. \end{aligned}$$

Thus in the interval  $(P_s, P')$  both parties are worse off than at  $P_s$ . If the bargaining outcome under the fixed technology falls in this interval then the removal of this "middle ground" through the adoption of the flexible technology means that the outcome must now fall to one side or the other of the interval. If the outcome falls to  $P_s$  or below, then the firm benefits; if it falls to  $P'$  or above, then the firm is hurt. Which will occur depends on the parameters of the problem.

In other words, the removal of this interval removes a region of possible compromise. In addition, outcomes in the interval  $(P', \infty)$  have been altered in a way that lowers the surplus over which the parties are bargaining. Again this makes compromise more difficult. Both these effects polarize bargaining by making it more of an "all or nothing" situation. The firm may benefit if it ends up with more ("all") or be hurt if it ends up with less ("nothing").

Note that the lost "middle ground" is independent of the bargaining rule. That is, the lost interval  $(P_s, P')$  can be determined without any reference to the bargaining rule. Thus, the results, particularly the result that flexibility can be detrimental, are robust to many different bargaining rules. (For example, instead of the Nash solution, we could have used an alternative bargaining rule suggested by Roth (1979, pp. 92–97) that maximizes the minimum of  $V(P)$  and  $W(P)$  after eliminating all Pareto-dominated outcomes.) In fact, we demonstrate in Appendix II that there need be no bargaining at all. In Appendix II we construct an example involving a monopolistic supplier and a price-taking downstream producer in which the adoption of flexibility by the downstream producer would be detrimental to it. Many different specifications of the objective functions led to the same basic conclusions, including those involving continuous quantities. In our example in Appendix II quantity can be any positive real number.

## II. Concluding Comments

It should be noted that, in our model, both advantages and disadvantages of flexi-

bility depend on an incentive-compatible threat that the flexibility confers on the firm. If this threat can be negotiated away then flexibility has neither positive nor negative value. In our example, if both price and quantity can be negotiated then the feasible set of outcomes under the flexible technology  $F$  is identical to that under the nonflexible technology  $A$ . Thus, in the presence of complete contracts flexibility will have zero value. In practice, however, firms make decisions over many different variables that affect both the firm and the input supplier. Firms and input suppliers may bargain over some of these variables. As long as there are variables that affect the input supplier but are excluded from bargaining there will be a potential for beneficial or detrimental (to the firm) flexibility. That is, the input supplier and the firm may bargain over many dimensions but as long as they do not bargain over all dimensions the results presented herein apply. (It is easy but messy to construct examples in higher dimensions.) It is this incompleteness that makes detrimental flexibility possible even in the light of a bargaining rule (such as Nash) that otherwise would be Pareto optimal.

Finally, note that we developed our model in a world of certainty in order to isolate theoretically the strategic aspects of flexibility from the risk-reducing (option) aspects. In a more realistic setting (one involving risk) a complete contract would have to be a complete contingent contract. Such contracts simply do not exist because of the high information and transaction costs. In the absence of such contracts, flexibility would confer both a strategic value and a value in risk reduction. We have shown in this paper that the strategic value of flexibility can, under some conditions, be negative.

#### APPENDIX I. FORMAL PROPOSITIONS

The model in the text is generalized and formalized below. We will use the generalized Nash solution to determine the outcome of bargaining. This involves maximizing  $\phi = V^{1-\delta}W^\delta$  for each of the three different technologies where  $\delta \in [0, 1]$  is bargaining power of the input supplier. (See

Roth, 1979, and Grout, 1984.) Note that

$$(A3) \quad \phi_A(P) = \begin{cases} (A - \alpha P)^{1-\delta} (\alpha P)^\delta & \text{if } 0 \leq P \leq A/\alpha \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_B(P) = \begin{cases} (B - \beta P)^{1-\delta} (\beta P)^\delta & \text{if } 0 \leq P \leq B/\beta \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_F(P) = \begin{cases} (A - \alpha P)^{1-\delta} (\alpha P)^\delta & \text{if } P > P_s \text{ and } 0 \leq P \leq A/\alpha \\ (B - \beta P)^{1-\delta} (\beta P)^\delta & \text{if } P > P_s \text{ and } 0 \leq P < B/\beta \\ 0 & \text{otherwise} \end{cases}$$

The input prices under each of these technologies is thus

$$P_A = \operatorname{argmax} \phi_A(P)$$

$$P_B = \operatorname{argmax} \phi_B(P)$$

and  $P_F = \operatorname{argmax} \phi_F(P).$

Then we can state the following propositions:

**PROPOSITION 1:** *If  $P_s \geq P_A$ , then  $P_F = P_A$  and flexibility has no value.*

**PROPOSITION 2:** *If  $P_s < P_A$ , then either a)  $P_F = P_s$  and flexibility is beneficial to the firm, or b)  $P_F = P_B$  and flexibility is detrimental to the firm.*

The following simple results are used in the proofs of the propositions:

R1:  $\phi_F(P) = \phi_A(P)$  if  $P \in [0, P_s]$

R2:  $\phi_F(P) = \phi_B(P)$  if  $P \in (P_s, \infty)$

R3:  $P_A = (\delta A)/\alpha$  uniquely maximizes  $\phi_A(P)$

R4:  $P_B = (\delta B)/\beta$  uniquely maximizes  $\phi_B(P)$

R5:  $\phi_A(P_A) > \phi_B(P_B)$

R6:  $\phi_A(P)$  is strictly increasing for  $P \in (0, P_A)$

R7:  $\phi_B(P)$  is strictly decreasing for  $P \in (P_B, B/\beta)$  and 0 for  $P \leq B/\beta$

R8:  $\phi_A(P_s) > \phi_B(P_s)$

The proofs of these simple results are straightforward and are omitted.

*Proof of Proposition 1.* If  $P_s \geq P_A$ , then  $\phi_F(P_A) = \phi_A(P_A)$ . It then follows from R1, R2, R3, R4, and R5 that  $P = P_A$  maximizes  $\phi_F(P)$ . Thus,  $P_F = P_A$  and mode  $A$  will be used. The firm gains nothing since fixed technology  $A$  would lead to an identical result.

*Proof of Proposition 2.* There are four possibilities given  $P_s < P_A$ : i)  $P_B \leq P_s$ ; ii)  $P_B > P_s$  and  $\phi_B(P_B) < \phi_A(P_s)$ ; iii)  $P_B > P_s$  and  $\phi_B(P_B) = \phi_A(P_s)$ ; and iv)  $P_B > P_s$  and  $\phi_B(P_B) > \phi_A(P_s)$ .

In the interval  $[0, P_s]$ ,  $\phi_F$  is maximized at  $P_s$  (by R1, R6, and  $P_s < P_A$ ). Now consider the interval  $(P_s, \infty)$ . In Case (i)  $\phi_F(P)$  is less than  $\phi_B(P_s)$  for all  $P$  in  $(P_s, \infty)$  (by R2, R7, and  $P_B \leq P_s$ ) and  $\phi_B(P_s) < \phi_F(P_s)$  (by R1 and R8). Thus,  $\phi_F$  is maximized at  $P_s$  or, in our notation,  $P_F = P_s$ . In Case (ii)  $P_B$  maximizes  $\phi_F$  in the interval  $(P_s, \infty)$  (by R2, R4, and  $P_B > P_s$ ), but  $\phi_F(P_B) < \phi_F(P_s)$  (by R1, R2,  $P_B > P_s$  and  $\phi_B(P_B) < \phi_A(P_s)$ ). Thus again,  $P_F = P_s$ . In both these cases flexibility is beneficial since  $P_s < P_A$  and

$$B - \beta P_s = A - \alpha P_s > A - \alpha P_A.$$

In Case iii) we get indeterminacy since  $\phi_F(P_s) = \phi_F(P_A)$ . This condition is a knife-edge that separates beneficial flexibility and detrimental flexibility.

In Case iv)  $P_B$  maximizes  $\phi_F$  in the interval  $(P_s, \infty)$  (by R2, R4, and  $P_B > P_s$ ) and  $\phi_F(P_B) > \phi_F(P_s)$  (by R1, R2,  $P_B > P_s$  and  $\phi_B(P_B) > \phi_A(P_s)$ ). Thus,  $P_F = P_B$ . In this case the firm is worse off since at  $P = P_B$ , the firm will use mode  $B$ . This is worse than the outcome of adopting the nonflexible technology  $A$  since

$$A - \alpha P_A = A(1 - \delta),$$

$$B - \beta P_B = B(1 - \delta),$$

$$\text{and} \quad A > B.$$

## APPENDIX II. A NUMERICAL EXAMPLE IN A DIFFERENT CONTEXT OF DETRIMENTAL FLEXIBILITY

Consider the case of an upstream monopolistic input supplier and a downstream price-taking producer. Suppose that the downstream firm has the choice of three technologies (two fixed,  $A$  and  $B$ , and one flexible,  $F$ ), each of which use the input. Suppose also that the profits under the technologies, given the input price  $P$ , are as follows:

### A. Profit Functions of Downstream Firm

$$\text{Technology } A: \pi = 3.5 - P$$

$$\text{Technology } B: \pi = 3.65 - 1.1P$$

$$\text{Technology } F: \pi = \begin{cases} 3.5 - P & \text{if } P \leq 1.5 \\ 3.65 - 1.1P & \text{if } P > 1.5 \end{cases}$$

Suppose the input demand of the downstream firm is given by the following demand functions:

### B. Input Demands of Downstream Firm

$$\text{Technology } A: Q = 4 - P$$

$$\text{Technology } B: Q = 2 - 0.5P$$

$$\text{Technology } F: Q = \begin{cases} 4 - P & \text{if } P \leq 1.5 \\ 2 - 0.5P & \text{if } P > 1.5 \end{cases}$$

Finally suppose that the upstream monopolist has the following cost function.

### C. Cost Function of Upstream Monopolist

$$\text{Total Cost} = (1/6)Q^3 - Q^2 + 3Q.$$

Simple profit-maximization results in the upstream monopolist's setting the following prices:

### D. Prices Set by the Monopolist

$$\text{Technology } A: P = 2.5858$$

$$\text{Technology } B: P = 3.1010$$

$$\text{Technology } F: P = 3.1010$$

This yields the following profits to the downstream producer:

E. *Profits to the Downstream Firm*

Technology A:  $\pi_s = 0.9428$

Technology B:  $\pi = 0.2323$

Technology F:  $\pi = -0.1042$

In this example it is clearly detrimental for the downstream firm to adopt a flexible manufacturing system, even if it is costless to install and to switch modes, and it is clear why this is so. The downstream producer's threat to switch to an input-conserving technology actually causes the input price to increase since demand is more inelastic in mode B.

#### REFERENCES

- Baldwin, Carliss, "Productivity and Labor Unions: An Application of the Theory of Self-Enforcing Contracts," *Journal of Business*, April 1983, 56, 155-85.
- Grout, Paul, "Investment and Wages in the Absence of Binding Contracts: A Nash Bargaining Approach," *Econometrica*, March 1984, 52, 449-60.
- Kulatilaka, Nalin, "The Value of Flexibility," MIT Energy Lab, Working Paper MIT-EL 86-014 WP, 1986, revised July 1987.
- Marks, Stephen, "A Strategic Model of Wage Bargaining and Investment Decisions," mimeo., School of Management, Boston University, September 1984.
- Mason, Scott and Merton, Robert, "The Role of Contingent Claims Analysis in Corporate Finance," in *Recent Advances in Corporate Finance*, E. I. Altman and M. G. Subrahmanyam, eds., Homewood, IL: Richard D. Irwin, 1985.
- McDonald, Robert and Siegel, Daniel, "Investment and the Valuation of firms When There Is an Option to Shutdown," *International Economic Review*, June 1985, 26, 331-49.
- \_\_\_\_\_ and \_\_\_\_\_, "The Value of Waiting to Invest," *Quarterly Journal of Economics*, November 1986, 101, 707-27.
- Nash, John, "The Bargaining Problem," *Econometrica*, April 1950, 28, 155-62.
- Roth, Alvin, *Axiomatic Models of Bargaining*, New York: Springer-Verlag, 1979.

# Accounting Rates of Return: Comment

By SHIMON AWERBUCH\*

In a recent contribution to this *Review*, Robert Anthony (1986) illustrates the differences between accounting and economic returns and shows that under certain conditions use of annuity depreciation serves to eliminate the so-called "accounting measurement error." Similar demonstrations (which follow from the seminal works of Harold Hotelling, 1925; Gabriel Preinreich, 1938; and Hector Anton, 1956) are given by Ezra Solomon, 1970; Zvi Bodie, 1982; William Beaver, 1981, pp. 77-81, and others.

Anthony seems to equate annuity with economic depreciation (1986, p. 245), although the two are not the same. He defines the general case as an "amortization schedule that 1) recovers the amount of investment over the life of the project, and 2) charges interest on the unrecovered amount" each year. Economic depreciation, which is the year-to-year change in the present value of the asset's future cash flows, will always meet Anthony's twin requirements, while annuity depreciation (AD) meets the conditions only when the cash flow profile (CFP) is level and unchanging over time. Perhaps the assumption of level cash flow explains why AD is "often called economic depreciation" (Anthony, 1986, p. 245) by accountants. General implementation of economic depreciation has already been found to be "totally impractical" (Franklin Fisher, 1984, p. 510) and not a "useful benchmark" (Edward McIntyre, 1977, p. 169). This comment therefore presumes that Anthony means annuity and not economic depreciation.

## I. Annuity Depreciation

AD, which, as Anthony notes, is analogous to the periodic schedule for allocating

interest and principal in simple mortgages, can be defined as

$$(1) \quad D_1 = R - i(PV)$$

for the first year, and

$$(2) \quad D_t = R - i \left( PV - \sum_{k=1}^{t-1} D_k \right)$$

for the general case (Anton, p. 122).  $D_1$  and  $D_t$  represent depreciation in year one and year  $t$ , respectively;  $R$  is the net annual (end of year) receipts;  $PV$  is the present value of those receipts,<sup>1</sup> and  $i$  is the interest or discount rate. The depreciation charges are also related through  $D_{t+1} = D_t(1+i)$ .

Anthony shows that under certain circumstances AD leads to correct accounting measurement of annual income and rate of return while straight-line depreciation (SLD) gives misleading results (1986, p. 245). Table 1 illustrates Anthony's point on the superiority of AD.<sup>2</sup> This demonstration must be interpreted with care, however, since similar examples can be constructed which give opposite results. Table 2, for example, shows a case in which AD (panel B) is "wrong" while SLD (panel A) clearly gives the correct 15 percent return each year, and meets Anthony's twin requirements. The declining CFP of Table 2 is more than an abstraction—it is anticipated by all rate-base regulated utilities, and may also be a reasonable expectation for *unregulated* firms given the widespread use of accelerated tax depreciation (Gerald Salamon, 1985, p. 499) and the presence of technological progress which drives down costs (Richard Bower, 1985, p. 10).

An infinite number of cash flow profiles exist for which AD produces accounting re-

\*Assistant Professor of Finance, College of Management Science, University of Lowell, Lowell, MA, 01854. The author is grateful for the comments of Richard S. Bower and Franklin M. Fisher and several anonymous referees.

<sup>1</sup>Although in practice, accountants simply use the initial investment (see Glenn Welsch, Charles Zlatkovich, and John White, 1976, p. 571).

<sup>2</sup>Without getting involved in his second argument regarding the cost of equity capital.

TABLE 1—ACCOUNTING INCOME AND RETURN: LEVEL CASH FLOW<sup>a</sup>

Year	Cash Flow <sup>b</sup>	Depreciation	Beginning Year		
			Assets	Net Income	ARR, in %
<b>A. Straight-Line Depreciation</b>					
0	\$-1,000.00				
1	298.32	\$200.00	\$1,000.00	\$98.32	9.8
2	298.32	200.00	800.00	98.32	12.3
3	298.32	200.00	600.00	98.32	16.4
4	298.32	200.00	400.00	98.32	24.6
5	298.32	200.00	200.00	98.32	49.2
<b>B. Annuity Depreciation</b>					
0	\$-1,000.00				
1	298.32	\$148.32	\$1,000.00	\$150.00	15.0
2	298.32	170.56	851.68	127.75	15.0
3	298.32	196.15	681.12	102.17	15.0
4	298.32	225.57	484.97	72.75	15.0
5	298.32	259.40	259.40	38.91	15.0

<sup>a</sup>This table illustrates Anthony's 1986 point regarding the superiority of AD.

<sup>b</sup>Net present value is zero at a discount rate of  $r=15$  percent.

TABLE 2—ACCOUNTING INCOME AND RETURN: REGULATED UTILITY CASH FLOW

Year	Cash Flow <sup>a</sup>	Depreciation	Beginning Year		
			Assets	Net Income	ARR, in %
<b>A. Straight-Line Depreciation</b>					
0	\$-1,000.00				
1	350.00	\$200.00	\$1,000.00	\$150.00	15.0
2	320.00	200.00	800.00	120.00	15.0
3	290.00	200.00	600.00	90.00	15.0
4	260.00	200.00	400.00	60.00	15.0
5	230.00	200.00	200.00	30.00	15.0
<b>B. Annuity Depreciation</b>					
0	\$-1,000.00				
1	350.00	\$148.32	\$1,000.00	\$201.68	20.2
2	320.00	170.56	851.68	149.44	17.6
3	290.00	196.15	681.12	93.85	19.3
4	260.00	225.40	484.97	34.43	7.1
5	230.00	259.90	259.40	29.40	-11.3

<sup>a</sup>Net present value is zero at a discount rate of  $r=15$  percent.

sults that are no better, and possibly worse, than the values obtained using ordinary SLD. Table 3, for example, gives the accounting income and returns, under SLD and AD, for the "Q Profile" used by Franklin Fisher and John McGowan, (F-M), 1983, Table 1. The profile has been adjusted in each case to yield an annual after-tax economic (IRR) return of 14.15 percent, the same return shown by F-M.<sup>3</sup>

<sup>3</sup>F-M's Table 1 erroneously reported the IRR as 15 percent.

F-M's Beginning-of-year ARR values (using sum-of-years' digits (SYD) depreciation) ranged from -5.3 percent to 50.3 percent, while the returns for SLD (panel A) range from -51.2 percent to 30.8 percent. Comparable returns under AD (panel B) range from -62.5 percent to 29.9 percent and are thus equally defective since they "vary substantially [and] never equal the economic rate of return" (F-M, 1983, p. 85).

Each of these isolated demonstrations, mine as well as Anthony's, takes advantage of one simple relationship: for every arbitrary depreciation schedule, be it SLD, AD,



TABLE 3—ACCOUNTING INCOME AND RETURN: THE "Q PROFILE"<sup>a</sup>

Year	Cash <sup>b</sup> Flow	Depreciation	Net Income	After-Tax Net Income	After-Tax Cash Flow	Beginning Year	
						Assets	ARR, in %
<b>A. Straight-Line Depreciation</b>							
1	\$24.2	\$16.67	\$7.6	\$4.2	\$20.8	\$100.0	4.2
2	45.9	16.67	29.2	16.1	32.7	83.3	19.3
3	54.0	16.67	37.4	20.5	37.2	66.7	30.8
4	42.2	16.67	25.5	14.0	30.7	50.0	28.1
5	21.0	16.67	4.3	2.4	19.1	33.3	7.2
6	8.1	16.67	-8.5	-8.5	8.1	16.7	-51.2
<b>B. Annuity Depreciation</b>							
1	\$25.1	\$11.7	\$13.4	\$7.4	\$19.1	\$100.0	7.4
2	47.5	13.3	34.2	18.8	32.1	88.3	21.3
3	56.0	15.2	40.8	22.4	37.6	75.0	29.9
4	43.7	17.4	26.3	14.5	31.9	59.8	24.2
5	21.8	19.8	2.0	1.1	20.9	42.4	2.5
6	8.4	22.6	-14.2	-14.2	8.4	22.6	-62.7

<sup>a</sup>See Fisher and McGowan (1983, Table 1).<sup>b</sup>The CF is scaled to produce an after-tax present value of \$100 at a discount rate of 14.15 percent, the same rate used by F-M.

TABLE 4—ACCOUNTING INCOME AND RETURN: VARIATIONS FROM LEVEL CASH FLOW

Year	Cash Flow <sup>a</sup>	Depreciation	Beginning Year		
			Assets	Net Income	ARR, in %
<b>A. Straight-Line Depreciation</b>					
	<i>b</i> = .95				
1	\$325.04	\$200.00	\$1,000.00	\$125.04	12.5
2	308.79	200.00	800.00	108.79	13.6
3	293.35	200.00	600.00	93.35	15.6
4	278.68	200.00	400.00	78.68	19.7
5	264.75	200.00	200.00	64.75	32.4
	<i>b</i> = .90				
1	\$353.89	\$200.00	\$1,000.00	\$153.89	15.4
2	318.50	200.00	800.00	118.50	14.8
3	286.65	200.00	600.00	86.65	14.4
4	257.99	200.00	400.00	57.99	14.5
5	232.19	200.00	200.00	32.19	16.1
<b>B. Annuity Depreciation</b>					
	<i>b</i> = .95				
1	\$325.04	\$148.32	\$1,000.00	\$176.72	17.7
2	308.79	170.56	851.68	138.23	16.2
3	293.35	196.15	681.12	97.20	14.3
4	278.68	225.57	484.97	53.11	11.0
5	264.75	259.40	259.40	5.35	2.1
	<i>b</i> = .90				
1	\$353.89	\$148.32	\$1,000.00	\$205.57	20.6
2	318.50	170.56	851.68	147.94	17.4
3	286.65	196.15	681.12	90.50	13.3
4	257.99	225.57	484.97	32.42	6.7
5	232.19	259.40	259.40	-27.21	-10.5

<sup>a</sup>The CF is reduced each year by  $(1 - b)$ ; net present value in all cases is zero at a discount rate of  $r = 15$  percent.

or SYD, there exists exactly one CFP which will yield an equivalence of accounting and economic returns (Thomas Stauffer, 1971). This somewhat obvious result merely says that for each profile there exists an economic depreciation schedule. AD is the economic depreciation in the case of level cash flows only, while SLD is the economic depreciation for the rate-base regulated CFP.

These CFP-depreciation relationships, however, are acutely sensitive and even slight changes significantly alter the outcome thus severely limiting the usefulness of Anthony's proposal that FASB mandate the use of AD over SLD. Table 4 illustrates that when CFP is "rotated" as little as 5 percent per year from the level profile shown by Anthony, AD is no better than SLD. On the basis of these results it seems difficult to justify the apparently widespread belief among economists that AD is superior.<sup>4</sup> Fully evaluated, the evidence suggests that use of AD is not likely to produce accounting measures that are better than those obtained with SLD since only when the cash flow is exactly level will AD yield the "correct" results.<sup>5</sup>

## II. The Needs of Financial Reporting

Anthony's prescription regarding depreciation has some other shortcomings which should make economists circumspect about undertaking the "task of persuading FASB" to universally adopt AD (1986, p. 246). It has long been recognized that income measures must be suited to the purpose (Myron Gordon, 1960, p. 608; John Hicks, 1969; Anton, p. 117) and can "quite rightly differ" depending on need (Kenneth Boulding, 1962, p. 45). The objectives of financial reporting deal with satisfying the needs of "investors and creditors" (Financial Accounting Standards Board, (FASB), 1978, p. 14), an orien-

tation that reflects a considerable shift away from the emphasis on income measurement (William Beaver, 1981, p. 4).

Given these objectives, there is a significant risk that use of AD will needlessly complicate the process of interpreting financial reports. Boulding, 1962, p. 54, has already cautioned against the use of "a complicated untruth" over "a simple untruth" in accounting and Beaver and Wayne Landsman, 1983, find that other *seemingly* sound practices, such as inflation adjusted reporting, do not enhance the information content. SLD therefore, simply makes it easier for investors to "know what the accountant's answer means," (Boulding, p. 53), which is important given Beaver and Joel Demski's demonstration (1979, p. 42) that income is poorly defined in the absence of perfect markets.

*Other Cash Flow Profiles.* The special needs of financial reporting notwithstanding, it may be possible to improve the economic usefulness of accounting reports by selecting the appropriate depreciation schedule a priori on the basis of the asset's *expected* CFP (Anton, p. 134). Although the problems raised by uncertainty<sup>6</sup> and the sensitivity of ARR to cash flow are not inconsequential, the results may be more informative since ARR will be constant (and equal to the IRR) whenever expectations are precisely met. Periodic deviation from expected results is thus clearly reflected<sup>7</sup> and depreciation becomes an empirical issue (Peter Luckett, 1984, p. 217; see also Alan Kraus and Ronald Huefner, 1972) with the hope that estimates improve through experience (Lawrence Gordon, 1974, p. 351). Though less than perfect, the approach provides a practical means of dealing with the problems of uncertainty and circularity that confound attempts to periodically determine true economic depreciation (Fisher, 1984, p. 510).

<sup>4</sup>The referees of an earlier version of this paper, for example, suggested that "It is difficult to see how the annuity method would have harmful effects in any setting," and that the income definition one gets using AD "is at least arguably superior."

<sup>5</sup>As one of those referees correctly observed: "One arbitrary depreciation method is no more informative than another."

<sup>6</sup>Without which the world would have little use for accounting; the reason you take the trouble to figure out what happened this year is that it may serve as a guide to what will happen next year.

<sup>7</sup>Assuming accountants correctly sift out price-level effects.

TABLE 5—CORRECT DEPRECIATION FOR THE "Q PROFILE"<sup>a</sup>

Year	Cash <sup>b</sup> Flow	Depreciation <sup>c</sup>	Net Income	After-Tax Net Income	After-Tax Cash Flow	Beginning Year	
						Assets	ARR, in %
1	\$34.88	\$9.15	\$25.73	\$14.15	\$23.30	\$100.00	14.14
2	45.52	22.14	23.38	12.86	35.00	90.85	14.15
3	45.06	27.38	17.68	9.72	37.10	68.71	14.15
4	33.49	22.85	10.63	5.85	28.70	41.33	14.15
5	17.54	12.78	4.75	2.62	15.40	18.48	14.15
6	7.16	5.69	1.47	0.81	6.50	5.69	14.15

<sup>a</sup>See Fisher and McGowan (1983, Table 1).<sup>b</sup>The CF is scaled to produce an after-tax present value of \$100 at a tax rate of 45 percent and a discount rate of 14.15 percent, the same rates used by F-M.<sup>c</sup>Using  $D_{t+1} = (1+i)D_t + (R_{t+1} - R_t)$ .

TABLE 6—CORRECT DEPRECIATION FOR A CYCLIC CASH FLOW

Year	Cash Flow <sup>a</sup>	Depreciation <sup>b</sup>	Beginning Year		
			Assets	Net Income	ARR, in %
1	\$100.00	\$61.56	\$256.26	\$38.44	15.00
2	10.00	-19.20	194.70	29.20	15.00
3	10.00	-22.09	213.90	32.09	15.00
4	100.00	64.60	235.99	35.40	15.00
5	10.00	-15.71	171.39	25.71	15.00
6	10.00	-18.06	187.09	28.06	15.00
7	100.00	69.23	205.16	30.77	15.00
8	10.00	-10.39	135.93	20.39	15.00
9	10.00	-11.95	146.32	21.95	15.00
10	100.00	76.26	158.27	23.74	15.00
11	10.00	-2.30	82.01	12.30	15.00
12	10.00	-2.65	84.31	12.65	15.00
13	100.00	86.96	86.96	13.04	15.00

<sup>a</sup>Net present value is zero at a discount rate of  $r=15$  percent.<sup>b</sup>Using  $D_{t+1} = (1+i)D_t + (R_{t+1} - R_t)$ .

It may therefore be useful to examine further appropriate depreciation charges for various CF profiles. For example, an investment in a stable, mature technology or product probably justifies the assumption of a level CFP coupled with AD (Table 1, panel B), whereas an investment in an asset with declining cash flows, such as an oil field, may warrant SLD (Table 2, panel A).<sup>8</sup> New products or technologies, on the other hand, may have cash flows similar to the "Q profile" (F-M, 1983, Table 1), for which ARR is quite different from IRR (IRR = 14.15 per-

cent) when SYD depreciation is used. The "correct" depreciation pattern (which meets Anthony's twin conditions) for the Q profile is considerably more *decelerated* than SYD (Table 5), reflecting the initially increasing CFP which results in larger depreciation charges in the second and third year during which significant portions of the asset's expectations are realized.

This effect is even more pronounced in the interesting case of an asset with a cyclic cash flow profile.<sup>9</sup> Table 6 shows the correct de-

<sup>8</sup>SLD is strictly correct only when  $D = (R_t - R_{t+1})/i$ .

<sup>9</sup>Real-life examples do not abound; two possibilities are a stand of timber that is completely harvested every few years and an Olympic sports complex such as the

preciation pattern for such an asset with  $ARR = IRR = 15$  percent. In this case depreciation is positive only in years during which revenue is received (years 1, 4, 7, 10, 13). In the intervening years the asset's value rises as the time for the next cash receipt draws nearer, hence resulting in negative depreciation.<sup>10</sup> Clearly such an asset is worth more directly before its quadrennial use (i.e., at the end of years 0, 3, 6, 9, 12) than one year earlier (i.e., at the end of years -1, 2, 5, 8, 11) even though it has aged chronologically. Accounting practice, however, has considerable difficulty with the notion of negative depreciation.<sup>11</sup>

### III. Conclusion

Accounting income concepts differ fundamentally from economic models as a consequence of numerous accounting decisions that are made in cost allocation, use of unadjusted historical figures, and arbitrary depreciation charges (Edgar Edwards and Philip Bell, 1961; Robert Sterling, 1970). Given the number of factors contributing to the problem it may indeed be "not reconcilable" (Yuji Ijiri, 1980, p. 54). In any event, it is unlikely that the "simple depreciation transformation" (McIntyre, 1977, p. 170) Anthony proposes will affect resolution. Any arbitrary depreciation pattern will yield acceptable results only for a very limited range of cash profiles. Anthony's proposal for universal use of AD will do little to generate "accounting reports that conform more closely to economic reality" (Anthony, 1986, p. 246). Such

an outcome requires changes in accounting procedures that are considerably more complex (and costly) as has been shown.

### REFERENCES

- Anthony, Robert N., "Accounting Rates of Return," *American Economic Review*, March 1986, 76, 244-46.
- Anton, Hector, A., "Depreciation, Cost Allocation and Investment Decisions," *Accounting Research*, April 1956, 7, 117-34.
- Beaver, William H., *Financial Reporting: An Accounting Revolution*, Englewood Cliffs: Prentice-Hall, 1981.
- \_\_\_\_\_ and Demski, Joel S., "The Nature of Income Measurement," *Accounting Review*, January 1979, 54, 38-46.
- \_\_\_\_\_ and Landsman, Wayne R., *Incremental Information Content of Statement 33 Disclosures*, Stamford: FASB, November 1983.
- Bierman, Harold Jr., "Depreciable Assets—Timing of Expense Recognition," *Accounting Review*, October 1961, 36, 613-18.
- Bodie, Zvi, "Compound Interest Depreciation in Capital Investment," *Harvard Business Review*, May-June 1982, 60, 58-60.
- Boulding, Kenneth, E., "Economics and Accounting: The Uncongenial Twins," in W. T. Baxter and Sidney Davidson, eds., *Studies in Accounting Theory*, Homewood: Irwin, 1962, 44-55.
- Bower, Richard S., "The Capital Recovery Question: An Overview," *Resources and Energy*, Vol. 7, Amsterdam: North-Holland, 1985, 7-42.
- Edwards, Edgar O. and Bell, Philip W., *Theory and Measurement of Business Income*, Berkeley: University of California Press, 1961.
- Fisher, Franklin M. and McGowan, John J., "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, 73, 82-97.
- \_\_\_\_\_, "On The Misuse of Accounting Rates of Return: Reply," *American Economic Review*, June 1984, 74, 509-17.
- Gordon, Lawrence A., "Accounting Rate of Return vs. Economic Rate of Return,"

facilities at Lake Placid, NY, which were built by a public-benefit corporation with the expectation of regular usage for recreation, training, and competitive events, coupled with peak usage every four years in conjunction with the winter Olympics. And while the income measure for a public authority may have a different significance than for an investor-owned corporation, economic depreciation is still a relevant concept.

<sup>10</sup>Sufficiently high inflation rates will yield negative depreciation for level or even declining real cash flows.

<sup>11</sup>F-M, 1983, p. 92, and Fisher, 1984, p. 510, observe that economic depreciation has not been adopted in practice because it can readily be negative, although Harold Bierman (1961, p. 616) shows how necessary accounting "write-ups" might be implemented.

- Journal of Business Finance and Accounting*, Spring 1974, 1, 343-56.
- Gordon, Myron J., "Scope and Method of Theory and Research in the Measurement of Income and Wealth," *Accounting Review*, October 1960, 35, 603-18.
- Hicks, John R., "The Measurement of Capital," *Proceedings of the 37th Session*, London, *Bulletin of the International Statistical Institute*, 1969, 43, 253-63; reprinted in Richard P. Brief, ed., *Depreciation and Capital Maintenance*, New York: Garland, 1984, 113-24.
- Hotelling, Harold, "A General Mathematical Theory of Depreciation," *Journal of the American Statistical Association*, September 1925, 20, 340-53.
- Ijiri, Yuji, "Recovery Rate and Cash Flow Accounting," *Financial Executive*, March 1980, 48, 54-56.
- Kraus, Alan and Huefner, Ronald J., "Cash-Flow Pattern and the Choice of a Depreciation Method," *Bell Journal of Economics and Management Science*, Spring 1972, 3, 316-34.
- Luckett, Peter F., "ARR vs. IRR: A Review & Analysis," *Journal of Business Finance and Accounting*, Summer 1984, 11, 213-31.
- McIntyre, Edward V., "Present Value Depreciation and the Disaggregation Problem," *Accounting Review*, January 1977, 52, 163-171.
- Preinreich, Gabriel A. D., "The Theory of Depreciation," *Econometrica*, July 1983, 6, 219-41.
- Salamon, Gerald L., "Accounting Rates of Return," *American Economic Review*, June 1985, 75, 495-504.
- Solomon, Ezra, "Alternative Rate of Return Concepts and Their Implications for Utility Regulation," *Bell Journal of Economics and Management Science*, Spring 1970, 1, 65-81.
- Stauffer, Thomas R., "The Measurement of Corporate Rates of Return," *Bell Journal of Economics and Management Science*, Autumn 1971, 2, 434-69.
- Sterling, Robert R., *Theory of the Measurement of Enterprise Income*, Lawrence: University of Kansas, 1970.
- Welsch, Glenn A., Zlatkovich, Charles T. and White, John A., *Intermediate Accounting*, 4th ed., Homewood: Irwin, 1976.
- Financial Accounting Standards Board, "Statement of Financial Accounting Concept No. 1: Objectives of Financial Reporting by Business Enterprises," Stamford, November 1978.

## Auditors' Report

February 23, 1988

Executive Committee  
The American Economic Association

We have examined the balance sheets of The American Economic Association as of December 31, 1987 and 1986, and the related statements of revenues and expenses, changes in general fund and restricted fund balances and changes in financial position for the years then ended. Our examinations were made in accordance with generally accepted auditing standards and, accordingly, included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the financial statements referred to above present fairly the financial position of The American Economic Association as of December 31, 1987 and 1986, its revenues and expenses and the changes in its financial position for the years then ended, in conformity with generally accepted accounting principles applied on a consistent basis.

Touche Ross & Co.  
Certified Public Accountants  
Nashville, Tennessee

## THE AMERICAN ECONOMIC ASSOCIATION BALANCE SHEETS, DECEMBER 31, 1987 AND 1986

	1987	1986
<b>Assets</b>		
CASH	\$ 850,226	\$ 576,067
INVESTMENTS, at market (Notes A and B)	4,759,969	4,835,255
ACCOUNTS RECEIVABLE, less allowance for doubtful accounts of \$590 (1987) and \$590 (1986)	31,075	127,807
INVENTORY OF <i>Index of Economic Articles</i> , at cost	205,958	137,148
PREPAID EXPENSES	18,830	15,698
OFFICE FURNITURE AND EQUIPMENT—at cost, less accumulated depreciation of \$61,175 (1987) and \$46,258 (1986)	66,520	65,742
	<u>\$5,932,578</u>	<u>\$5,757,717</u>
<b>Liabilities and Fund Balances</b>		
ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	\$ 510,849	\$ 368,121
DEFERRED REVENUE (Note A):		
Life membership dues	39,192	41,814
Other membership dues	615,242	573,718
Subscriptions	524,592	471,646
<i>Job Openings for Economists</i>	21,468	20,850
	<u>1,200,494</u>	<u>1,108,028</u>
ACCRUAL FOR DIRECTORY (Note A)	208,313	138,313
FUND BALANCES:		
General	3,898,027	3,982,063
Net worth	3,898,027	3,982,063
Restricted	114,895	161,192
Total Fund Balances	<u>4,012,922</u>	<u>4,143,255</u>
	<u>\$5,932,578</u>	<u>\$5,757,717</u>

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF REVENUES AND EXPENSES  
FOR THE YEARS ENDED DECEMBER 31, 1987 AND 1986

	1987	1986
<b>REVENUES FROM DUES AND ACTIVITIES:</b>		
Membership dues and subscriptions	\$ 909,310	\$ 870,387
Nonmember subscriptions	664,605	649,607
<i>Job Openings for Economists</i> subscriptions	33,536	32,499
Advertising	131,999	130,025
Sale of <i>Index of Economic Articles</i>	89,617	43,247
Sale of copies, republications, and handbooks	36,516	40,256
Sale of mailing list	49,153	51,127
Annual meeting	58,770	36,948
Sundry (Exhibit I)	81,322	72,514
Operating Revenues	<u>2,054,828</u>	<u>1,926,610</u>
<b>PUBLICATION EXPENSES:</b>		
<i>American Economic Review</i>	614,199	612,751
<i>Journal of Economic Literature</i>	812,990	767,648
Directory publication (Note A)	70,000	70,000
<i>Job Openings for Economists</i>	52,943	55,351
<i>Index of Economic Articles</i>	57,054	30,615
<i>Journal of Economic Perspectives</i>	250,685	75,002
	<u>\$1,857,871</u>	<u>1,611,367</u>
<b>OPERATING AND ADMINISTRATIVE EXPENSES:</b>		
General and administrative:		
Salaries	178,532	172,271
Rent	18,360	16,768
Other (Exhibit II)	189,086	202,862
Committee	44,474	45,286
Annual meeting	4,331	6,644
	<u>434,783</u>	<u>443,831</u>
Operating Expenses	<u>2,292,654</u>	<u>2,055,198</u>
Operating Deficit	(237,826)	(128,588)
INVESTMENT INCOME RECOGNIZED (Note B)	<u>270,564</u>	<u>248,027</u>
REVENUES IN EXCESS OF EXPENSES	<u>\$ 32,738</u>	<u>\$ 119,439</u>

See notes to financial statements.



## THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN GENERAL FUND BALANCE

	Total	Operations	Market Value Adjustments
Balance at January 1, 1986	\$3,217,603	\$2,152,914	\$1,064,689
Add change in market value of investments	645,021	-	645,021
Add revenues in excess of expenses	119,439	119,439	-
Balance at December 31, 1986	<u>3,982,063</u>	<u>2,272,353</u>	<u>1,709,710</u>
Add change in market value of investments	(116,774)	-	(116,774)
Add revenues in excess of expenses	<u>32,738</u>	<u>32,738</u>	<u>-</u>
Balance at December 31, 1987	<u>\$3,898,027</u>	<u>\$2,305,091</u>	<u>\$1,592,936</u>

See notes to financial statements.

## THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN RESTRICTED FUND BALANCE

	Balance at January 1	Receipts	Disburse- ments	Balance at December 31
YEAR ENDED DECEMBER 31, 1986:				
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	\$105,447	\$171,918	\$140,836	\$136,529
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	-	-	5,000
The Rockefeller Foundation Grant for minority students applying for graduate work in economics	(14,087)	125,001	95,144	15,770
Sundry	2,693	4,100	2,900	3,893
	<u>\$ 99,053</u>	<u>\$301,019</u>	<u>\$238,880</u>	<u>\$161,192</u>
YEAR ENDED DECEMBER 31, 1987:				
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	\$136,529	\$140,782	\$170,309	\$107,002
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	-	-	5,000
The Rockefeller Foundation Grant for minority students applying for graduate work in economics	15,770	122,021	137,791	-
Sundry	3,893	100	1,100	2,893
	<u>\$161,192</u>	<u>\$262,903</u>	<u>\$309,200</u>	<u>\$114,895</u>

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN FINANCIAL POSITION  
FOR THE YEARS ENDED DECEMBER 31, 1987 AND 1986

	1987	1986
Cash, beginning of year	\$576,067	\$632,751
Cash Used in Operations:		
Revenues in excess of expenses	32,738	119,439
Items that did not (provide) use cash:		
Depreciation	14,254	10,418
Directory publication (Note A)	70,000	70,000
Investment income recognized (Note A)	(270,564)	(248,027)
Cash used in operations	(153,572)	(48,170)
INCREASE (DECREASE) IN CASH DUE TO CHANGES IN:		
Investments	75,286	(507,470)
Accounts receivable	96,732	(18,874)
Inventory of <i>Index of Economic Articles</i>	(68,810)	(36,288)
Prepaid expenses	(3,132)	(2,852)
Office furniture and equipment	(15,032)	(14,860)
Accounts payable and accrued liabilities	142,728	(160,722)
Deferred revenue	92,466	54,175
Accrual for directory	-	(445)
Restricted funds	(46,297)	62,139
General fund, market value adjustments	153,790	616,683
Cash, end of year	<u>\$850,226</u>	<u>\$576,067</u>

See notes to financial statements.

## Notes to Financial Statements

### A. Summary of Significant Accounting Policies

*Investments* are accounted for on a market value basis. The investment income recognized is modified to reflect only the Association's approximate historical average rate of return, which is currently 5%. Investment income represents 5% of the total cash and market value of investments at the beginning of the year. The change in market value of investments and dividends and interest earned net of investment income recognized is recorded directly to the general fund.

*The Accrual for directory* results because every three to five years the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was most recently published in 1985 and distributed at no cost to the membership. In order to properly match the publishing cost of this directory with revenue from membership dues, the Association provided \$70,000 in 1987 and \$70,000 in 1986 for estimated publishing costs which will reduce actual directory expenses in the year of publication.

*Deferred revenue* represents income from membership dues and subscriptions to the various periodicals of the Association which are deferred when received. These amounts are then recognized as income following the distribution of the specified publications to the members and subscribers of the Association. Income from life membership dues is recognized over the estimated average life of these members.

The American Economic Association files its federal income tax return as an educational organization, substantially exempt from income tax under Section 501(c) (3) of the Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists. The Association has been determined to be an organization which is not a private foundation.

## B. Investments and Investment Income

Investments consist of:

	December 31, 1987		December 31, 1986	
	Cost	Market	Cost	Market
Government obligations, bonds, and commercial paper	\$ 607,196	\$ 656,868	\$ 980,658	\$1,105,080
Corporate stocks and mutual funds	4,252,090	4,103,101	3,216,323	3,730,175
	<u>\$4,859,286</u>	<u>\$4,759,969</u>	<u>\$4,197,481</u>	<u>\$4,835,255</u>

Investment income recognized consist of:

	Year Ended December 31	
	1987	1986
Government obligations, bonds, and commercial paper—interest	\$123,780	\$132,454
Corporate stocks and mutual funds—cash dividends	263,952	341,239
Change in market value	(233,942)	142,990
Transfer to general fund, net	116,774	(368,656)
Investment income recognized, net	<u>\$270,564</u>	<u>\$248,027</u>

## C. Retirement Annuity Plan

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was approximately \$40,000 and \$36,000 for 1987 and 1986, respectively.

## D. Ratio of Net Worth to Expenses

The ratio of net worth at December 31, 1987 to 1988 budgeted expenses is 1.47 and the ratio of net worth at December 31, 1986 to actual 1987 expenses is 1.74.

**EXHIBIT I—THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF  
SUNDRY REVENUES FOR THE YEARS ENDED  
DECEMBER 31, 1987 AND 1986**

	1987	1986
<i>AER</i> submission fees	\$35,685	\$27,312
Royalties	30,546	26,982
<i>CSWEP</i> membership dues	14,415	13,938
Donations	492	374
Miscellaneous income	348	3,509
Permission to reprint	—	341
Foreign postage	(164)	58
	<u>\$81,322</u>	<u>\$72,514</u>

**EXHIBIT II—THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF OTHER  
GENERAL AND ADMINISTRATIVE EXPENSES FOR THE YEARS ENDED  
DECEMBER 31, 1987 AND 1986**

	1987	1986
Dues and subscriptions	\$ 41,002	\$ 61,600
Mailing list file maintenance	26,468	27,073
Investment counsel and custodian fees	20,578	17,334
Accounting and legal	17,790	19,003
Postage	17,379	17,882
Periodic mailing expenses	15,523	14,117
Depreciation (straight-line method)	14,254	10,418
Office supplies	10,241	13,460
Bank charges	8,598	8,440
President and president-elect expenses	6,040	6,398
Telephone	5,601	5,196
Insurance and miscellaneous	5,273	4,791
Travel and entertainment	339	300
Uncollectible receivables	—	(3,150)
	<u>\$189,086</u>	<u>\$202,862</u>

# HANDBOOKS IN ECONOMICS

Editors:

**Kenneth J. Arrow**, Stanford University, Palo Alto, CA, USA

**Michael D. Intriligator**, University of California, Los Angeles, CA, USA

The **Handbooks in Economics** series continues

to provide the various branches of economics with handbooks which are definitive reference sources, suitable for use by professional researchers, advanced graduate students, or by those seeking a teaching supplement.

With contributions from leading researchers, each **Handbook** presents an accurate, self-contained survey of the current state of the topic under examination. These surveys summarize the most recent discussions in journals, and elucidate new developments. Although original material is also included, the main aim of this series is the provision of comprehensive and accessible surveys.

The **Handbooks** are indispensable reference works which belong in every professional collection, and form ideal supplementary reading for graduate economics students on advanced courses.

## Survey of the series

**1. Handbook of Mathematical Economics**  
(in 4 volumes)

Edited by K.J. Arrow and M.D. Intriligator  
*Vol. IV to be published in 1989*

**2. Handbook of Econometrics**

(in 3 volumes)

Edited by Z. Griliches and M.D. Intriligator

**3. Handbook of International Economics**

(in 2 volumes)

Edited by R.W. Jones and P.B. Kenen

**4. Handbook of Public Economics**

(in 2 volumes)

Edited by A. Auerbach and M. Feldstein

*Vol. 2 to be published in 1987/88*

**5. Handbook of Labor Economics**

(in 2 volumes)

Edited by O.C. Ashenfelter and R. Layard

**6. Handbook of Natural Resource and Energy Economics** (in 3 volumes)

Edited by A.V. Kneese and J.L. Sweeney

*Vol. 3 to be published in 1987/88*

**7. Handbook of Regional and Urban Economics**

(in 2 volumes)

Edited by E.S. Mills and P. Nijkamp

*Vol. 2 to be published in 1987*

**8. Handbook of Monetary Economics**

Edited by B. Friedman and F.H. Hahn

*To be published in 1987/88*

**9. Handbook of Development Economics**

Edited by H. Chenery and T.N. Srinivasan

*To be published in 1987/88*

**10. Handbook of Industrial Organization**

Edited by R. Schmalensee and R. Willig

*To be published in 1987/88*

**11. Handbook of Game Theory with Economic Applications**

Edited by R. Aumann and S. Hart

*To be published in 1989*

**12. Handbook of the Economics of Finance**

Edited by R. Merton and M.S. Scholes

*To be published in 1989*

Please mail to:

**North-Holland**

(A Division of Elsevier Science Publishers BV.)

Attn.: C. Fennes

P.O. Box 1991, 1000 BZ Amsterdam

The Netherlands

For customers in the USA and Canada:

**Elsevier Science Publishing Co., Inc.**

P.O. Box 1663, Grand Central Station

New York, NY 10163, U.S.A.

**North-Holland**

☐ Please send me your brochure on the **Handbooks in Economics** series

Name: \_\_\_\_\_

Address: \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

**AER**

NH/ECON/BK/2360

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

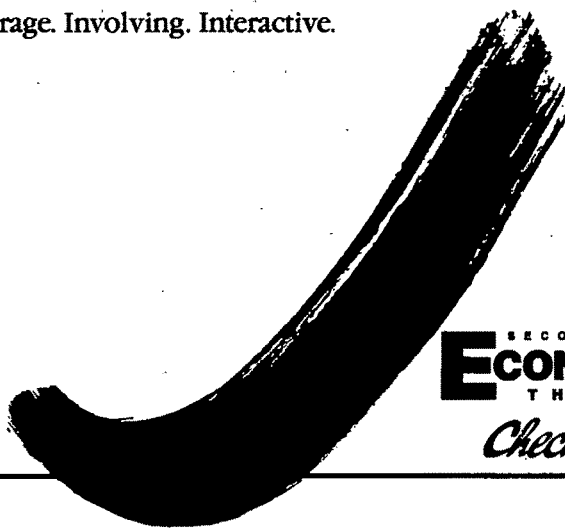
## Supply your students with the knowledge they'll need tomorrow. In a style they demand today.

With Allen R. Thompson's **Economics, Second Edition**, your students get a solid base in economics that helps them understand the world around them. That means focused coverage of the principles. Within the framework of contemporary theory.

Your students learn to think analytically...like economists, not technicians. And they truly enjoy learning, because up-to-date examples and applications keep them involved.

What's more, Thompson's **Economics, Second Edition** gives them broad coverage of policy issues—not superficial coverage of a lot of topics, or in-depth studies of only a few. Plus the book has a timely international flavor, through many examples from the foreign sector.

Solid coverage. Involving. Interactive.



SECOND EDITION  
**ECONOMICS**  
THOMPSON

*Check it out!*

A D D I S O N • W E S L E Y • E C O N O M I C S

## More top-notch texts...

N E W • F O R • 1 9 8 9

### **Macroeconomics**

John M. Barron, *Purdue University*, Mark A. Loewenstein, *Virginia Polytechnic & State University*, and Gerald J. Lynch, *Purdue University*

Incorporating the microfoundations of macroeconomics, the authors provide a balanced, modern approach to the study of macroeconomics. They discuss a variety of macroeconomic models in the context of aggregate demand and aggregate supply analysis. In addition a fully integrated treatment of financial markets is included, and all analytical chapters incorporate open economy analysis. (13623)

### **The Practice of Econometrics**

Ernst Berndt, *Massachusetts Institute of Technology*

This applied econometrics text gives advanced undergraduates and graduate students hands-on experience with estimation and inference. Can be used as a supplementary text for any course in econometric theory. Or as a principal text for a one semester course in applied econometrics.

### **An Introduction to Economic Reasoning**

William D. Rohlfs Jr., *Drury College*

Rohlfs teaches economics as a way of thinking by clearly developing a simple analytic approach and by applying it to numerous applications and current policy issues. *An Introduction to Economic Reasoning* will engage your students... with features such as...a fully developed discussion of supply and demand, in-text readings, a chapter on pricing, and an integrated study guide. (15743)

A L S O • A V A I L A B L E

### **Econometrics: An Introduction**

T. Dudley Wallace, *Duke University*, and J. Lew Silver, *Emory University* (09896)

### **Walrasian Microeconomics: An Introduction to the Economic Theory of Market Behavior**

Donald W. Katzner, *University of Massachusetts* (10461)

### **Economics: A Tool for Understanding Society, Third Edition**

Tom Riddell, *Smith College*, Jean Shackelford, *Bucknell University*, and Steve Stamos, *Bucknell University* (06368)



**Addison-Wesley Publishing Company**

Reading, Massachusetts 01867 • (617) 944-3700

# BLACKWELL

## **Trade, Development and the World Economy**

Selected Essays of Carlos Diaz-Alejandro

**Edited by ANDRES VELASCO**

This volume consists of selected papers in international economics and development written over the past twenty-five years by Carlos Diaz-Alejandro, covering trade policies for development, foreign investment and multinational corporations, North-South relations, and less developed country perspectives on stabilization policies and the international monetary system.

416 pages, **\$75.00** (0 631 15687 9)

## **Unemployment, Hysteresis and the Natural Rate Hypothesis**

**Edited by ROD CROSS, assisted by A. Chawluk and M. Malek**

This important new volume is the first to investigate the relevance of hysteresis effects to the analysis of unemployment.

The papers deal with the significance of hysteresis effects for natural rate theory; the characteristics of the unemployed; the microeconomic rationale for hysteresis effects in labour markets; the explanation of aggregate unemployment; and the policy implications.

448 pages, **\$75.00** (0 631 15688 7)

## **Regulation of the Firm and Natural Monopoly**

**MICHAEL WATERSON**

Provides a comprehensive analysis and assessment of the purpose and effects of the main forms of regulation from the standpoint of industrial economics. The focus of the study is on industries where market conduct is potentially substantially different from competitive.

176 pages, **\$49.95** (0 631 14007 7)

## **Unified Theory of Estimation & Inference for Nonlinear Dynamic Models**

**A. RONALD GALLANT and HALBERT WHITE**

Building on the great advances that have been made by statisticians and econometricians in the last forty years, this study presents the first unified theory of estimation and inference applicable to a wide variety of econometric estimators of the parameters of time-dependant heterogenous economic phenomena.

164 pages, **\$34.95** (0 631 15765 4)



**Basil Blackwell**

108 Cowley Road, Oxford OX4 1JF  
Suite 1503, 432 Park Avenue South, New York NY 10016.  
Toll-free ordering: 1-800-638-3030 (USA)



# AEA sponsored Group Life Insurance for you and your family— at attractive rates!

The AEA Group Life Insurance Plan can help provide valuable supplementary protection—at attractive rates—for eligible members and their dependents.

Because AEA participates in a large Insurance Trust which includes other scientific and technical organizations, the low cost may be even further reduced by premium credits. In the past nine years, insured members received credits on their April 1 semiannual payment notices averaging 40% of their annual premium contributions. (These credits are based on the amount paid during the previous policy year ending September 30.) Of course future premium credits, and their amounts, cannot be promised or guaranteed.

Now may be a good time for you to re-evaluate your present coverage and look into AEA Life Insurance. Just fill out and return the coupon for more details at no obligation.

**Administrator, AEA Group Insurance Program**  
1255 23rd Street, N.W.  
Washington, D.C. 20037

J-4

Please send me more information about the AEA Life Insurance Plan.

Name \_\_\_\_\_ Age \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Or—call today Toll-Free 800-424-9883  
(Washington, DC area, call 296-8030)

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

R O B E R T S. D A N I E L L.

**PINDYCK / RUBINFELD**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

UNIVERSITY OF CALIFORNIA, BERKELEY

# **MICROECONOMICS**

**T H E O R Y   A T   W O R K**

AVAILABLE FOR FALL ADOPTION, AUGUST 1988  
HAVE A LOOK • GIVE US A CALL AT 1-800-428-3750

**M A C M I L L A N**

COLLEGE DIVISION, 866 THIRD AVE., NEW YORK, NY 10022

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# INTERNATIONAL ECONOMIC JOURNAL

VOLUME 2

SPRING 1988

NUMBER 1

The Effects of Government Intervention in a Dynamic Model of the Spot  
and Forward Exchange Markets

Jay H. Levin

Strategic Investment Determining Time of Entry

Hyung Bae

Fiscal Policy and the Current Account: What Do Capital Controls Do?

J. M. Vinals and J. T. Cuddington

Test of Rationality of Adaptive Expectations in the German Hyperinflation

Inchul Noh

Domestic Adjustment Policies and External Economic Shocks

Parvez Hasan

Is the Observed Intra-Industry Trade a Statistical Artifact?

Tecksung Kwon

Successful Adjustment in a Multi-Sectoral Economy

Joshua Aizenman

Nipponized Confucian Ethos or Incentive-Compatible Institutional Design:  
Notes on Morishima, "Why Has Japan Succeeded?"

Henry Wan, Jr.

*International Economic Journal is published quarterly in English by the Korea International Economic Association. The personal subscription price of a volume (which includes postage) is \$30.00 per year. The institutional subscription price is \$40.00 per year. Cheques should be made payable to the International Economic Journal and sent directly to the Managing Editor. (Volume 1 is available.)*

*MANUSCRIPTS SUBMITTED in Asia for publication should be sent to the Managing Editor, Professor Wontack Hong, Department of International Economics, College of Social Sciences, Seoul University, Seoul 151-742, Korea. Manuscripts from countries outside Asia should be sent to the Co-Editor, Professor Young Chin Kim, Department of Economics, Northern Illinois University, Dekalb, Illinois 60115, U.S.A.*

## ECONOMETRIC SOFTWARE FROM TSP INTL

**Now available for PCs and mainframe computers**

**TSP Version 4.1:** with Probit, Logit, Tobit, sample selection, general maximum likelihood estimation, and robust standard errors for all procedures. A complete programming language for econometricians and data analysts in use at over 1000 sites worldwide, it includes regression, nonlinear simultaneous equation estimation, time series models, and many other features.

The PC version is interactive and identical to the mainframe version. Databanks and the saving and restoring of workspaces are now supported. PC TSP requires 512K RAM, a math chip (8087 or 80287), hard disk recommended.

**For mainframes only**

**RATS Version 2.0:** identical to the popular PC version. We are the authorized distributor of the mainframe version.

For more info, write or call (415) 326-1927  
TSP International • PO Box 61015 • Palo Alto, CA 94306

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# Looking at

## **THE RELATIVE INEFFICIENCY OF QUOTAS**

**James E. Anderson**

James Anderson demonstrates here that in most reasonable circumstances quotas are an inferior trade policy relative to import tariffs. Presenting substantive new work on tariffs and quotas in imperfect competition, he provides a better understanding of quotas and protection policies.

\$25.00

## **INVESTMENT CHOICES IN INDUSTRY**

**Constance E. Helfat**

Portfolio theory until now has been used mainly to allocate investments in financial assets such as stocks and bonds; here Constance Helfat lays the groundwork for increasingly broader and more varied application of portfolio theory to the making of operating decisions in firms.

\$25.00

## **FISCAL POLICIES AND THE WORLD ECONOMY**

**An Intertemporal Approach**

**Jacob A. Frenkel and Assaf Razin**

*"Fiscal Policies and the World Economy has all the ingredients to be the most significant book in international finance in the last 15 years."*

—Jeremy Greenwood, University of Western Ontario

\$24.00

## **PRIVATIZATION**

**An Economic Analysis**

**John Vickers and George Yarrow**

This comprehensive analysis of the British privatization program offers insights into recent policies on privatization, competition, and regulation in a country that has by far the greatest experience with this growing world-wide phenomenon.

\$39.95

# Our World Economy

## **THE THEORY OF INDUSTRIAL ORGANIZATION**

**Jean Tirole**

"This book will fill a tremendous void in the textbook market for advanced undergraduate and graduate level courses in industrial organization and applied microeconomics. The strength of Tirole's work is his masterful synthesis of analytical development and intuitive discussion."—John P. Borin, Professor of Economics, Wesleyan University

\$35.00

*Original in Paperback*

## **ECONOMIC EFFECTS OF THE GOVERNMENT BUDGET**

**edited by Elhanan Helpman, Assaf Razin, and Efraim Sadka**

The original contributions in this book analyze all of a government's budget's components—expenditures, revenues, the deficit—with a special emphasis on recently important issues, such as intergenerational transfers of debt and declines in corporate tax revenues.

\$15.00 paper (\$37.50 cloth)

## **TAX POLICY AND THE ECONOMY, VOLUME 2**

**edited by Lawrence H. Summers**

Second in a series of annual publications on tax policy and the economy initiated by the National Bureau of Economic Research and designed to convey research results in a way that is accessible to a wide body of lawyers, policymakers, and business people involved in formulating tax policy.

\$12.95 paper (\$25.95 cloth)

## **THE EVOLUTION OF CENTRAL BANKS**

**Charles Goodhart**

Are central banks a necessary intervention in today's free-market banking system? Many economists have challenged the value of central banks, but Charles Goodhart provides an authoritative and original approach in applying the considerable insights of history to concerns of current policymaking.

\$11.95 paper (\$22.50 cloth)

## **The MIT Press**

55 Hayward Street, Cambridge, MA 02142

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

BILATERALISM,  
MULTILATERALISM  
AND CANADA IN  
U.S. TRADE POLICY

WILLIAM DIEBOLD, JR.

A Council on Foreign Relations Book

## Bilateralism, Multilateralism, and Canada in U.S. Trade Policy

**William Diebold, Jr.,**  
Editor

*A Council on Foreign  
Relations book*

Five experts from the United States, Canada, and Mexico explore the critical issues raised by the U.S.-Canada bilateral free trade agreement and make recommendations for U.S. policy.

**April 1988** **224 pages**  
**0-88730-287-4** **cloth, \$16.95**

## Competing for Control *America's Stake in Microelectronics*

**Michael Borrus**

*A BRIE book*

Using the core industry of microelectronics as a case study, Michael Borrus examines the Japanese challenge to American industry and shows how to restore our competitive position.

**June 1988** **312 pages**  
**0-88730-306-4** **cloth, \$32.00**

## Cooperation and Competition in the Global Economy: *Issues and Strategies*

**Antonio Furino, Editor**

*An IC<sup>2</sup> book*

**June 1988** **256 pages**  
**0-88730-307-2** **cloth, \$35.00**

Approaching cooperation and competition as complementary rather than mutually exclusive objectives, prominent scholars, policymakers, and business leaders discuss American industry's role in the international marketplace.

## Keeping Pace *U.S. Policies and Global Economic Change*

**John Yochelson, Editor**

*A CSIS book*

Economists and policymakers examine the changing U.S. role in the world economy in light of the stunning technological advances that are transforming the global power equation.

**June 1988** **352 pages**  
**0-88730-252-1** **cloth, \$34.95**

## The Impact of Technological Change on Employment and Economic Growth

**Richard M. Cyert and David C. Mowery,**  
Editors

*A National Academy of Sciences book*

This book assesses employment effects of technological change — from the impact of new technologies on firm size to their implications for the distribution of income and employment in the U.S. economy.

**July 1988** **480 pages**  
**0-88730-290-4** **cloth, \$39.95**

**YES! Please send me:**

- ☐ **Bilateralism and Canada** (6612667) \$16.95
- ☐ **Competing for Control** (6612758) \$32.00
- ☐ **Cooperation and Competition** (6612857) \$35.00
- ☐ **Keeping Pace** (6612279) \$34.95
- ☐ **Impact of Technological Change** (6612709) \$39.95

☐ Payment enclosed ☐ Bill me  
Charge my ☐ MC ☐ VISA ☐ AMX

Card no. \_\_\_\_\_ Exp. date \_\_\_\_\_

Signature \_\_\_\_\_

Send to: \_\_\_\_\_

Zip \_\_\_\_\_

**BALLINGER**  
**PUBLISHING COMPANY**

*Harper & Row*  
Order Department  
2350 Virginia Avenue  
Hagerstown, MD 21740

**(800) 638-3030**

My state sales tax \$ \_\_\_\_\_

Postage/handling  
(\$2.75/order)\* \$ \_\_\_\_\_

**\*Prepaid orders are postage free!**

**TOTAL \$ \_\_\_\_\_**

Prices subject to change. All orders subject to credit approval. U.S. funds only. If you order by phone, tell the operator your order code is **AAER688**

## Integrates SPREADSHEET + STATISTICS + FORECASTING + GRAPHICS and is Easy to Use.

- ySTAT is the only full-fledged spreadsheet statistical program - a fast, powerful, easy-to-use stand-alone program.
- Menu-driven format as in Lotus 1-2-3 - offers many functions that are not available even in 1-2-3, such as vector formulas, lagged variables, dummy variables, moving averages, and interpolation of missing values.
- You can read, enter, edit, copy, move, and sort the data as well as group and partition the data for analysis right on the spreadsheet. A total of 40 function keys for one-keystroke operations.
- Ten statistical function keys: to tally sums and means; subtract variable means or seasonal means; standardize the data; aggregate seasonal data; generate seasonal indicators; perform moving averages, linear trend forecasts, and so on.
- ySTAT reads 1-2-3 and Symphony worksheet files directly, and any free-format or fixed-format textfile generated by other PC programs or downloaded from mainframe. The data can be numeric, alphabetic, including missing values.
- Mean, median, quartiles, skewness, mode, t-test, frequency distribution, correlation, crosstabs, analysis of variance.
- Multiple regression (OLS), standardized regression, weighted least squares, two-stage least squares, polynomial regression, and Cochrane-Orcutt method. Including diagnostic residual analysis output (see sample at right).
- Pooling of cross-section and time-series data.
- Two-sample difference of means test, nonparametric tests (Kolmogorov-Smirnov test, Wald-Wolfowitz runs test, Mann-Whitney test, Wilcoxon matched-pairs signed-ranks test) and Spearman's and Kendall's rank-order correlations.
- Time-series models: autoregressive model, moving average model, and Box-Jenkins model.
- Nonseasonal forecasts: linear trend, quadratic trend, polynomial model, sinusoidal model; simple, double and triple exponential smoothings.
- Seasonal forecasts: exponential smoothings with seasonal indicators or with trigonometric functions; Winters' additive and multiplicative procedures.
- LOGIT, PROBIT, TOBIT, and Weighted Probit regressions: these models are twice as fast as Gauss program and faster than other programs and are much easier to use.

High-resolution color charts: the actual and forecast values are graphically displayed as a line chart and also as an XY chart, including a regression line and its 95% confidence interval.

OLS -- DEPENDENT VARIABLE: Consumpt

RIGHT-HAND VARIABLE	ESTIMATED COEFFICIENT	STANDARD ERROR	T-STATISTIC	PROB.
1 Profits	0.19234381	0.09121	T= 2.11327	0.049
2 P-1	0.00844898	0.03063	T= 0.27588	0.385
3 Wm	0.796218750	0.03994	T= 19.93342	0.000
4 Constant	16.23660272	1.30270	T= 12.46382	0.000

SAMPLE SIZE( 1 to 21) = 21

DF=17

SUM OF SQUARED RESIDUALS =	17.87948
VARIANCE (MSE) =	1.051732
STANDARD ERROR (ROOT MSE) =	1.025540
ADJUSTED R-SQUARED =	0.981008
F-STATISTIC( 3, 17) =	292.707585 (p=0.0004)
SUM OF RESIDUALS =	-0.000000
DURBIN-WATSON STATISTIC =	1.567474

Analysis of Variance		Source	SUM SQ	DF	MEAN SQ
Due to Regression			923.550	3	307.850
Residual			17.879	17	1.052
Total			941.430	20	47.071

RESIDUAL ANALYSIS -Mean: -0.000 Adj. RSE: 0.546 Mean Abs. % Err: 1.263

AUTOCORRELATION

LAG	COEF	T-VAL
1	0.181	0.83
2	-0.059	-0.27
3	-0.033	-0.15
4	-0.157	-0.69
5	-0.411	-1.91

Ljung-Box statistic (chi-square 4 DF): 6.678 (p=0.1527)

PARTIAL AUTOCORRELATION

LAG	COEF	T-VAL
1	0.181	0.83
2	-0.035	-0.44
3	-0.004	-0.02
4	-0.162	-0.74
5	-0.378	-1.73

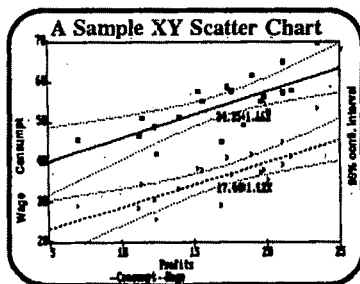
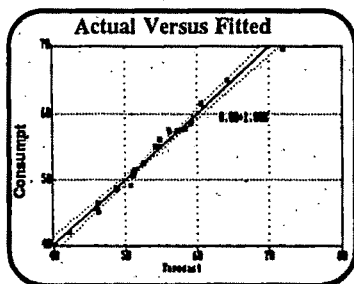
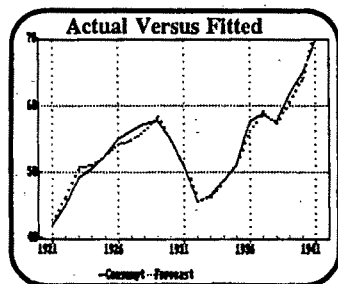
ACTUAL versus FITTED VALUES and RESIDUALS

SEQ	Actual	Fitted	min=41.900	71.973-max	Error
1	42.80	42.22			-0.584
2	45.00	46.23			1.230
3	49.20	50.77			1.566
4	50.60	51.09			0.494
5	52.60	52.58			0.008
6	55.10	54.23			0.869
7	56.20	54.86			1.334
8	57.30	56.35			0.955
9	57.80	58.39			-0.589
10	58.00	54.72			0.262
11	50.80	51.17			-0.370
12	45.60	45.82			-0.222
13	46.50	46.18			-0.322
14	48.70	49.76			-1.058
15	51.30	51.33			-0.035
16	57.70	56.08			1.616
17	56.70	58.70			-2.030
18	57.50	57.29			0.210
19	61.60	60.61			0.989
20	65.00	64.22			0.785
21	69.70	71.67			-2.173

LIST OF OUTLIERS - STUDENTIZED RESIDUALS GREATER THAN 1.80

SEQ	Actual	Fitted	Residual	Std. Err	Studentized Probability	Residuals	DF=17
3	49.20	50.77	-1.566	0.867	-1.619	0.124	
16	57.70	56.08	1.616	0.539	1.687	0.110	
21	69.70	71.67	-2.173	0.580	-2.787	0.013	

(Note: this sample output shows some of the options for regression diagnostics.)



### MING TELECOMPUTING INC.

23 Oak Meadow Road, P.O. Box 101, Lincoln Center, MA 01773, U.S.A.

(617) 259-0391

Order: ( ) ySTAT for \$395. ( ) ySTAT/Medical for \$395.  
 ( ) Trial Disk and information/sample output for \$5.  
 ( ) Information/sample output.  
 System: ( ) IBM PC ( ) XT ( ) AT ( ) System 2 Model  
 ( ) Other system  
 Options: ( ) with 8087 or 80287 or 80387 math co-processor.  
 ( ) IBM CGA ( ) EGA ( ) Hercules Mono. Card  
 Type of printer or plotter

Payment: ( ) Check or money order.

( ) U.S. university or governmental purchase order.

( ) Visa. ( ) MC. Card No. \_\_\_\_\_

Name \_\_\_\_\_ Expir. date \_\_\_\_/\_\_\_\_/\_\_\_\_

Address \_\_\_\_\_

Telephone: ( ) \_\_\_\_\_

IBM is a registered trademark of International Business Machines Corp.; 1-2-3 of Lotus Development Corp.; Gauss of Aptech Systems, Inc.

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

## AMERICAN ECONOMIC ASSOCIATION

### 1988 ANNUAL MEMBERSHIP RATES

#### Membership includes:

— a subscription to *The American Economic Review* (quarterly) plus *Papers and Proceedings*, the *Journal of Economic Literature* (quarterly) and the *Journal of Economic Perspectives* (quarterly).

- Regular members with annual incomes of \$30,000 or less ..... \$38.50
- Regular members with annual incomes above \$30,000 but no more than \$40,000 ..... \$46.20
- Regular members with annual incomes above \$40,000 ..... \$53.90
- Junior members (available to registered students for three years only).

Student status must be certified by your major professor or school registrar ..... \$19.25

- In Countries other than the U.S.A., Add \$16.00 to cover postage.
- Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association) ..... \$7.70

Please enter my subscription for the following period:

☐ Jan.-Dec.      ☐ April-March      ☐ July-June      ☐ Oct.-Sept.

First Name and Initial	Last Name	Suffix
Address Line 1		
Address Line 2		
City		
State or Country	Zip/Postal Code	

**MAJOR FIELDS (TWO ONLY)**  
LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM JEL, "Classification System for Books."

Please type or print information above. Please pay with a check or money order payable in United States Dollars. Canadian and foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Please note: It is the policy of the Association, not to refund membership payments.

Endorsed by (AEA member) \_\_\_\_\_

#### Below for Junior Members Only

I certify that the person named above is enrolled as a student at \_\_\_\_\_

\_\_\_\_\_  
Authorized Signature

PLEASE SEND WITH PAYMENT TO:

**AMERICAN ECONOMIC ASSOCIATION**  
1313 21ST AVENUE SOUTH, SUITE 809  
NASHVILLE, TENNESSEE 37212-2786  
U.S.A.



# INDEX OF ECONOMIC ARTICLES

prepared by

*The Journal of Economic Literature*  
of the  
*American Economic Association*

- ✓ Each volume in the **Index** lists articles in major economic journals and in collective volumes published during a specific year.
- ✓ No other single reference source covers as many articles classified in economic categories as the **Index**.

**Index** volumes XI-XXII covering 1969-1980 are available at \$60.00 each.

The following two part **Index** volumes are now ready for delivery at \$90.00 per set:

Volume	Year
XXIII	1981
XXIV	1982
XXV	1983
XXVI	1984

Note: AEA members  
are entitled to a  
30% discount.

*an  
indispensable  
tool for...*

ECONOMISTS  
REFERENCE LIBRARIANS  
RESEARCHERS  
TEACHERS  
STUDENTS  
AUTHORS

Payment required in advance. Prices include shipping charges; allow 4-6 weeks for delivery. Please send your check or money order (net of applicable discount) payable in United States dollars drawn on a United States bank to:

American Economic Assn. - **Index**  
P.O. Box 20111  
Nashville, TN 37202

Address inquiries or other correspondence to:  
Journal of Economic Literature, P.O. Box 7320  
Oakland Station, Pittsburgh, PA 15213.

# THE ECONOMICS INSTITUTE

Gateway to Successful  
Master's and Doctoral Degree Studies  
in Economics, Agricultural Economics, Business and Administration

*Since its establishment in 1958 under the sponsorship of the American Economic Association, the Institute has provided specialized preparatory training and orientation to over 6,000 students from over 120 countries en route to over 320 universities in the United States and other English-speaking countries.*

## PROGRAM FEATURES

- Six to forty-six week individualized programs.
- Intensive and comprehensive review or supplementary preparatory training in:  
**English, Computer Usage, Economic Theory, Mathematics, Statistics, Accounting, Management, Finance and Marketing.**
- Standardized test preparation (TOEFL, GRE, GMAT).
- Orientation to U.S. campus and community life.
- University placement assistance services.
- Certificate and Diploma Programs for short-term professional trainees.

## END RESULTS for

- **Associated universities:** Improved admissions procedures and a source of additional high-quality foreign students.
- **Foreign students:** Greater accessibility to U.S. universities.
- **Sponsoring organizations:** Reduced total costs.
- **And for all three:** Better grade performance records in degree programs and more rapid completion of degree requirements.

## ORGANIZATION

### **AEA Policy and Advisory Board:**

Edwin S. Mills, Northwestern University, Chairman

Lance E. Davis, California Institute of Technology

Koichi Hamada, Yale University

Joseph Hawlicek, Ohio State University

Teh-wei Hu, University of California, Berkeley

Samuel A. Morley, Vanderbilt University

Ray Marshall, University of Texas at Austin

Stefan H. Robock, Columbia University

### **Director**

Wyn F. Owen, University of Colorado

**The Economics Institute, 1030 13th Street, Boulder, Colorado, 80302, USA;**  
**Telephone: (303)492-3000; Telex: 450385 ECONINST BDR; FAX: (303)492-3006**

## INTERNATIONAL MONETARY FUND STAFF PAPERS

*Staff Papers* contains studies by the Fund's staff on international monetary and financial problems. Among the articles included in the March 1988 issue are:

*Modeling and Testing Ricardian Equivalence: A Survey* .....

by Leonardo Leiderman and Mario I. Blejer

*Exchange Rates and the Term Structure of Interest Rates* .....

by James M. Boughton

*Exchange Rate Variability and the Slowdown in Growth of International Trade* .....

by Paul De Grauwe

*Currency Substitution in Egypt and the Yemen Arab Republic: A Comparative Quantitative Analysis* .....

by Mohamed A. El-Erian

*Government Spending, the Real Interest Rate, and the Behavior of Liquidity-Constrained Consumers in Developing Countries* .....

by Nicola Rossi

*The Macroeconomic Effects of Tax Reform in the United States* .....

by Owen Evans and Lloyd Kenward

### Subscription rates:

US\$18.00 per volume, US\$5.00 for a single issue. Special rates for university libraries, faculty, and students: US\$9.00 per volume.

Available from: Publication Services • Box E-391

International Monetary Fund • 700 19th Street, N.W.

Washington, D.C. 20431, U.S.A. • Telephone (202) 623-7430

## FEDERAL RESERVE BANK OF NEW YORK RESEARCH FUNCTION



Research positions available for experienced economists with strong publication record in macroeconomics, money, finance or international economics. Temporary and permanent appointments available. Salaries commensurate with experience and level of achievement. Equal Opportunity Employer. For more information contact:

Charles A. Pigott

Recruiting Officer

Federal Reserve Bank of New York, Room 937

New York, New York 10045

(212) 720-6321

# GENERAL EQUILIBRIUM MODELS FOR MICRO-COMPUTERS

## GEMODEL USA

### A large-scale general equilibrium model of the U.S. economy

The package is completely menu-driven and accomplishes every job from data entry through consistency checks, calibration and policy changes to solution and report printing.

The model can be used to analyze the effects of changes in tax policy, product demands, technology, productivity etc. on incomes, outputs, tax revenue and foreign trade.

GEMODEL USA is for policy analysis and research in government, university and industry.

A special purpose of the model is to simulate value-added taxes and the separate tax policies of two levels of government:  
Federal  
State/Local

GEMODEL USA has a phenomenal number-crunching capability. It accepts the 85-industry, U.S. input-output tables prior to aggregation of industries, consumption categories, and households. The level of detail carried by the model is larger than that commonly found in the general equilibrium literature.

The package is easy to run and is perhaps the only software that produces simulation results on your own desk within seconds. It is available in two versions: Professional and Academic.

PRICE:  
Academic Version US\$ 2,200

---

SYSTEM REQUIREMENTS: IBM-PC or compatible, 256K RAM, 2 drives, DOS 2.1 or better.

---

## GEMODEL 2.0

Solves open or closed, two or three-industry small open economy models employing two or three factors. Industries can have constant or diminishing returns to scale. Admits one to nineteen household groups with or without income/leisure choice. Solves with a wide array of factor, commodity and income taxes.

PRICE: US\$ 395.00

Ideal to enrich courses in Price Theory, Trade and Finance by solving models within seconds at 48 different levels of complexity with 18 to 176 equations.

## GESTATS

Calibrates GEMODEL 2.0 parameters to data and elasticity assumptions. Checks data consistency.

PRICE: US\$ 195.00

## GEREPORT

Compares GEMODEL 2.0 simulation results with the base case or with other simulation results. Computes measures of welfare change and marginal costs of taxes.

PRICE: US\$ 95.00

## GEDATA

18 data files for exercises in price theory, trade and finance using GEMODEL 2.0

PRICE: US\$ 25.00

For information and to order, write to:

**DIA Agency Inc.**, 1879 Kingsdale Ave., Ottawa, Ontario, K1T 1H9 CANADA  
DEMONSTRATOR diskettes available at \$25.



# THE CRITICS ACCLAIM THE NEW PALGRAVE A DICTIONARY OF ECONOMICS

Edited by  
John Eatwell, Murray Milgate, Peter Newman

"The result is an unparalleled store of information...*The New Palgrave* is a testimony to the capacity of economics to illustrate the world!"

ROBERT HEILBRONER  
NEW YORK REVIEW OF BOOKS

"...far surpasses any other reference work on economics."

WASHINGTON POST

"*The New Palgrave* offers today's trained economists an inexhaustible supply of instruction and provocation...here are hundreds of articles that convey the fascination and importance of the subject."

THE ECONOMIST

"...an event of great importance to all economists...a *tour de force*."

RICHARD STONE

"To attempt such an undertaking was audacious, to have carried it out is astonishing, and to have done so with so remarkable a list of authors is a tribute to the intelligence and diligence of the editors."

ALAN S. BLINDER

"...an indispensable reference tool...the topics are exhaustive."

KENNETH J. ARROW

"The list of contributors reads like a who's who in economics. The range of topics is breathtaking...it is overwhelmingly impressive."

MARTIN L. WEITZMAN

"...a remarkably ambitious and extraordinarily useful enterprise."

J.K. GALBRAITH

"It will undoubtedly become the classic for this century."

BOOKLIST/REFERENCE BOOKS BULLETIN

PRICE RISES TO \$750 ON JULY 31, 1988.

PUBLICATION DATE: JANUARY 1988

\* 4 Volumes

\* 927 Contributors \* 4,194 Pages

\* 1,261 Subject Entries \* 655 Biographies

\* 4.3 Million Words \* 4,000 Cross References

ORDERS RECEIVED BEFORE  
THAT DATE:

\$595 prepaid orders

\$650 billed orders

(institutions and corporations only)

For further information, or to order, please call 800 221 2123. In NY call collect 212 481 1334.

Or write to the publisher, Stockton Press, 15 East 26th St., New York, NY 10010.

We accept VISA, MASTERCARD, DINERS, AMERICAN EXPRESS

## **INDIRECT TAXATION IN DEVELOPING ECONOMIES**

*Second Edition*

**John F. Due**

"[Due] has drawn upon his unsurpassed general knowledge of the field. . . . It is a synthesis which distills the essence of the issues."—James A. Maxwell, *Journal of Economic Literature* (review of the first edition)

Now in a completely rewritten second edition, this remains the definitive treatment of the subject by the foremost authority in the field.

*The Johns Hopkins Studies in Development*  
Vernon W. Ruttan and T. Paul Schultz, Consulting Editors

**\$28.50**

## **FACTS AND FIGURES ON GOVERNMENT FINANCE**

*1988-1989 Edition*

**The Tax Foundation**

Published regularly since 1941—and now available from Johns Hopkins—*Facts and Figures on Government Finance* brings together data on public finance at all levels of government. Combining material from hundreds of sources—including official government documents and private sources, many of them inaccessible or out of print—the book furnishes the statistical information necessary to any rational and meaningful debate on issues in public finance.

**\$24.95 paperback \$30.00 hardcover**

## **NONLINEAR PREFERENCE AND UTILITY THEORY**

**Peter C. Fishburn**

This is the first major systematic treatment of the rapidly developing field of nonlinear preference and utility structures. Fishburn describes the traditional utility theories, identifies their shortcomings, and reviews alternate theories that overcome these shortcomings.

*Johns Hopkins Series in Mathematical Sciences*

**\$47.50**

## **WORLD POPULATION PROJECTIONS, 1987-88**

*Short- and Long-Term Estimates*

**K. C. Zachariah and My T. Vu**

This expanded edition of World Bank population estimates and projections contains the most up-to-date assessment of demographic prospects around the world. Detailed profiles compiled from the most recent population data available are presented for every country and for groups of countries classified by geographical region and income level.

*Published for the World Bank*

**\$52.95**



**THE JOHNS HOPKINS UNIVERSITY PRESS**

701 West 40th Street, Suite 275, Baltimore, Maryland 21211

## New Books in Economics from \_\_\_\_\_

**Greenwood Press,** 

**Praeger Publishers,**



**and Quorum Books** 

### **PROGRESS FOR FOOD OR FOOD FOR PROGRESS?**

The Political Economy of Agricultural Growth  
and Development  
by **Folke Dovring**

A comparative analysis of agricultural growth  
and development around the world.

ISBN 0-275-92904-3. \$49.95 Praeger

### **A SLIPPERY SLOPE**

The Long Road to the Breakup of AT&T  
by **Fred W. Henck** and **Bernard Stassburg**

Presents, for the first time, a complete history of  
the events that led to the breakup of the Bell  
System.

ISBN 0-313-26025-7. \$37.95 Greenwood

### **TEXTILES IN TRANSITION**

Technology, Wages, and Industry Relocation in  
the U.S. Textile Industry, 1880-1930

by **Nancy Frances Kane**

Contributes a valuable new approach to the  
study of relocation and wage differentials in the  
U.S. textile industry with analysis centering on  
the reasons for the timing of relocation of  
American textile production from the Northeast  
to the South and the simultaneous pattern of  
wage convergence between the two regions.

ISBN 0-313-25529-6. \$37.95 Greenwood

### **A REVOLUTION IN ARREARS**

The Development Crisis in Latin America  
by **Leland M. Wooton**

Postulates that economic development in Latin  
America is suffering from an unfinished social  
revolution.

ISBN 0-275-92689-3. \$39.95 Praeger

### **US-CHINA TRADE**

Problems and Prospects

edited by **Eugene K. Lawson**

"Lawson has drawn together some of the ablest  
people working on trade with China, all of  
them with one kind or another of hands-on  
experience...highly recommended to anyone  
interested in trade with China."

—Dwight H. Perkins, Director,  
Harvard Institute for International  
Development

ISBN 0-275-92494-7. \$49.95 Praeger

### **ENTREPRENEURSHIP AND PUBLIC POLICY**

Can Government Stimulate Business Startups?  
by **Benjamin W. Mokry**

Considers whether it makes sense for govern-  
ment to devise policies that will nurture small  
business ventures without a more thorough  
exploration of the dynamics of entrepreneurship  
and the possible impact of government initiatives.

ISBN 0-89930-239-4. \$35.95 Quorum

### **EDUCATION INCORPORATED**

School-Business Cooperation for Economic  
Growth

edited by the **Northeast-Midwest Institute.**

Identifies ways education and business can  
work together to build a strong economy.

ISBN 0-89930-282-3. \$39.95 Quorum

**Greenwood Press**

**Praeger Publishers**

**Quorum Books**

**Divisions and Imprint of Greenwood Press, Inc.**

88 Post Road West

P.O. Box 5007

Westport, CT 06881

(203) 226-3571

*New Titles from*  
**Cambridge University Press**

**The Foreign Exchange Market**

*Theory and Econometric Evidence*  
**Richard Baillie and Patrick MacMahon**

Provides an integrated approach to recent developments in the understanding of foreign exchange markets. It covers the theory of efficient markets as developed in finance, and the models used to explain the movements of exchange rates in macroeconomics.

About \$37.50

**Debt Problems of Eastern Europe**  
**Iliana Zloch-Christy**

Analyzes the causes and consequences of the massive East European debt that began in the 1970s and the flaws of the centrally-planned economies that led to the crisis as well as the lack of effective structural adjustment.

**Soviet and East European Studies**  
\$37.50

**The European Monetary System**

**Francesco Giavazzi, Stefano Micossi, and Marcus Miller, Editors**

Contains the papers and proceedings of a conference organized by the Centre for Economic Policy Research in cooperation with the Banca d'Italia and the Centro Interuniversitario di Studi Teorici per la Politica Economica.

About \$49.50

**High Public Debt**

*The Italian Experience*

**Francesco Giavazzi and Luigi Spaventa**

This book, based on a conference organized by the Centre for Economic Policy Research, looks at how the public and private sectors in Italy have found ways to adapt to high and rapidly increasing levels of debt, and examines the longer-term costs of their strategies.

About \$44.50

**Predictability in Science and Society**

**Sir John Mason, P. Mathias, and J.H. Westcott, Editors**

Surveys the uses and limitations of predictive modeling. The problems and difficulties of making accurate predictions are discussed in this volume by drawing examples from Newtonian dynamics, economics, meteorology, Marxism, and political studies.

Paper \$24.95

**The Growth and Efficiency of Government Spending**  
**Malcolm Levitt and Michael Joyce**

This book analyzes the growth and pattern of public spending in Britain since the 1960s and provides alternative estimates for the 1990s. Against this background the authors outline the problems associated with the interpretation of the concept of output in the public services.

**National Institute of Economic and Social Research Occasional Paper XLI**  
\$39.50



## **Dynamic Econometric Modeling**

*Proceedings of the Third International Symposium in Economic Theory and Econometrics*

**William A. Barnett, Ernst R. Berndt, and Halbert White, Editors**

Brings together presentations of some of the fundamental new research that has begun to appear in the areas of dynamic structural modeling, nonlinear structural modeling, time series modeling, nonparametric inference, and chaotic attractor inference.

**International Symposia in Economic Theory and Econometrics**  
\$49.50

*Now in paperback...*

## **International Capital Movements**

**Charles P. Kindleberger**

This study of international capital movements looks at their historical role in the financing of trade and their dramatically increased role in the world economy in recent years.

*About \$9.95*

## **Achieving Industrialization in East Asia**

**Helen Hughes, Editor**

This book examines the economic success of the newly industrializing and near-industrializing economies of East Asia. The distinguished group of authors covers a range of topics in a comparative perspective, and identifies lessons of concern to economic, political, and social questions throughout the developing world.

*About \$42.50*

*Announcing a new series...*

## **Structural Analysis in the Social Sciences**

**Mark Granovetter, Editor**

This new series is designed to bring together under a single rubric social scientific research undertaken from a structural perspective.

## **Intercompany Relations**

*The Structural Analysis of Business*

**Mark S. Mizruchi and Michael Schwartz, Editors**

This volume constitutes the first compilation of work by leading international scholars who have adopted a structural approach to the study of business.

**Structural Analysis in the Social Sciences 1**  
\$39.50

## **Social Structures**

*A Network Approach*

**Barry Wellman and S.D. Berkowitz, Editors**

This collection of original articles demonstrates the case for structural analysis through a variety of studies addressing key sociological questions, such as social mobility, business enterprise, the activity of nations, and revolutionary change.

**Structural Analysis in the Social Sciences 2**  
Cloth \$65.00 Paper \$22.95

At bookstores or order from

**Cambridge University Press**

32 East 57th Street, NY, NY 10022

Cambridge toll-free numbers for orders only:  
800-872-7423, outside NY State. 800-227-0247,  
NY State only. MasterCard and Visa accepted.

# THE ECONOMIC CHALLENGE OF PERESTROIKA

**ABEL  
AGANBEGYAN**

Chief Economic  
Adviser to **MIKHAIL  
GORBACHEV**

**Indiana  
University  
Press**

**is proud to  
announce  
a major  
publishing  
event**

"... perestroika is revolutionary in character and its full realization will alter and renew our whole society. . . . although it is proceeding slowly, with difficulty, and many mistakes have already been made along the way, with more probably yet to come, nonetheless, as Gorbachev has said, there is nowhere for us to retreat. We must move forward."

—from THE ECONOMIC CHALLENGE OF PERESTROIKA

Abel Aganbegyan, the most significant economist in contemporary Soviet society and Mikhail Gorbachev's chief architect of *perestroika*, has written a uniquely authoritative, comprehensive, and candid analysis of the current economic reconstruction in the USSR. *The Guardian* has called Aganbegyan the "big wheel of *perestroika*" and said that "he remains disarmingly frank about the problems which beset the Soviet economy." This book was specially commissioned for the Western reader to clarify the ideas, processes, and implications of *perestroika* for the future.

In the new series, *Second World*, edited by Teodor Shanin of the University of Manchester.

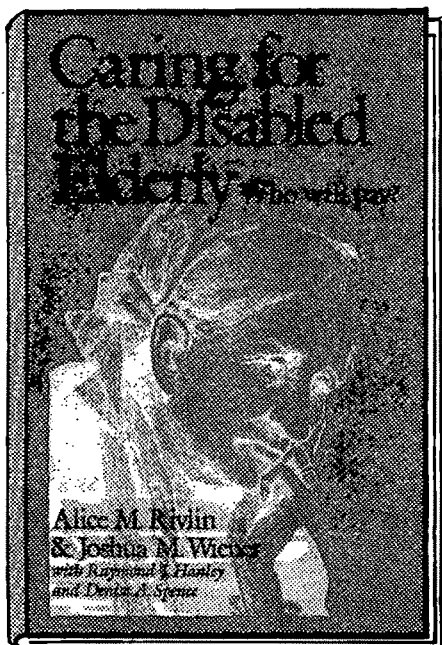


**AT BOOKSTORES EVERYWHERE**

**\$18.95**

**INDIANA  
UNIVERSITY PRESS**

Tenth and Morton Streets, Bloomington, Indiana 47405



## **Caring for the Disabled Elderly**

### **Who Will Pay?**

*Alice M. Rivlin and  
Joshua M. Wiener*

The disabled elderly and their families confront a delivery and financing system that is fragmented, inefficient, and expensive. This major new book addresses the fundamental policy questions that must be faced if we are to offer better long-term care in the future.

Cloth \$29.95/Paper \$11.95

## **Liability**

### **Perspectives and Policy**

*Robert E. Litan and  
Clifford M. Winston, Editors*

Liability experts provide clear and insightful descriptions of the major factors that have contributed to the "insurance crisis" and set forth a methodological framework for evaluating the debate over the current liability system.

Cloth \$28.95/Paper \$10.95

## **Innovation and the Productivity Crisis**

*Martin Neil Baily and  
Alok K. Chakrabarti*

The collapse of U.S. productivity growth since the late 1960s has been the most severe and persistent of recent economic problems. This volume reviews the extent of the growth slowdown, evaluates several contributing factors, and suggests strategies for improvement.

Cloth \$22.95/Paper \$8.95

# **BROOKINGS**

**The Brookings Institution**  
1775 Massachusetts Avenue NW  
Washington, D.C. 20036  
(202) 797-6258

# ANALYSIS & INSIGHTS

Studies in economics from Chicago



## THE FIRM, THE MARKET, AND THE LAW

R. H. COASE

Until now, R. H. Coase's most important papers have not been readily available. This collection brings together the classic articles "The Nature of the Firm" and "The Problem of Social Cost" with four other papers that clarify and extend Coase's arguments and a new introductory essay.

Cloth \$29.95 232 pages  
1 line drawing

## MERGERS AND ACQUISITIONS

Edited by Alan J. Auerbach

*Mergers and Acquisitions* surveys, in a nontechnical format, some of the important issues arising from the recent takeover boom, offering a valuable resource for making appropriate policy decisions about the costs and benefits of corporate takeovers.

Cloth \$17.95 120 pages  
8 line drawings  
An NBER Project Report

## FISCAL FEDERALISM

*Quantitative Studies*

Edited by Harvey S. Rosen

This volume examines the interrelation between state, local, and federal governments and the roles they play in U.S. fiscal policy, considering such issues as trends in fiscal centralization and the role of intergovernmental transfers.

Cloth \$39.95 (est.) 256 pages  
4 line drawings 9 maps  
An NBER Project Report

## INTERNATIONAL ECONOMIC COOPERATION

Edited by Martin Feldstein

*International Economic Cooperation* makes available the proceedings of an NBER conference organized to investigate efforts to coordinate economic policy among the developed nations. The volume includes background papers prepared by Stanley Fischer, Richard C. Marston, J. David Richardson, and Jeffrey D. Sachs; personal statements by individuals prominent in government and business, including W. Michael Blumenthal, Alan Greenspan, Edmund T. Pratt, Jr., Helmut Schmidt, and Robert S. Strauss; and summaries of the discussion that followed presentations.

Paper \$17.95 (est.) 345 pages  
15 line drawings  
Library cloth edition \$50.00 (est.)  
An NBER Conference Report

## THE UNITED STATES IN THE WORLD ECONOMY

Edited by Martin Feldstein

"This is an extraordinarily comprehensive and valuable collection that deserves a wide readership. *The United States in the World Economy* will appeal to specialists in both macroeconomics and international economics as a valuable reference work. The lucidity of the papers makes the book accessible to a much wider audience as well, those readers who are economically literate though not necessarily professional economists." — Stanley Fischer, Massachusetts Institute of Technology

Paper \$24.95 712 pages  
40 line drawings  
Library cloth edition \$75.00  
An NBER Conference Report

## PENSIONS IN THE U.S. ECONOMY

Edited by Zvi Bodie, John B. Shoven, and David A. Wise

*Pensions in the U.S. Economy* reports on retirement saving of individuals and the saving that results from corporate funding of pension plans as well as on aspects of the plans themselves from the employee's point of view.

Cloth \$28.00 208 pages  
18 line drawings  
An NBER Project Report

THE UNIVERSITY OF CHICAGO PRESS

5801 South Ellis Avenue Chicago, IL 60637

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# North-Holland

## New Books in Economics

### Productivity and U.S. Economic Growth

By D.W. Jorgenson, F.M. Gollop and B. Fraumeni

Contributions to Economic Analysis, 169

1988 About 504 pages  
Price: Dfl. 135.00  
ISBN 0-444-70353-5

*Published jointly with and distributed outside Europe and the UK by Harvard University Press, USA*

One of the most important features of the book is the way in which it successfully integrates the theory of producer behavior with the indexing and measurement of production growth. The authors present startling new findings showing that less than one-fourth of overall growth is attributable to advances in productivity.

### Statistical Data Bank Systems

**Socio-Economic Database and Model Building in Japan**

Edited by K. Uno and S. Shishido

1988 x + 362 pages  
Price: US \$92.00/Dfl. 175.00  
ISBN 0-444-70397-7

This book reports on the latest developments in statistical database and quantitative analysis in Japan. The articles are the outcome of the Project MUSE (Multiple-Use Socio-Economic) Data Bank. The purpose of the research was to promote interdisciplinary research regarding the feasibility of a statistical data bank in the social science field in response to increasing social needs.

### Spatial Analysis and Planning under Imprecision

By Y. Leung

Studies in Regional Science & Urban Economics, 17

1988 xviii + 376 pages  
Price: US \$86.75/Dfl. 165.00  
ISBN 0-444-70390-X

The book deals with complexity, imprecision, human valuation, and uncertainty in spatial analysis and planning, providing a systematic exposure of a new philosophical and theoretical foundation for spatial analysis and planning under imprecision.

### Macroeconomic Modelling

By S.G. Hall and S.G.B. Henry

Contributions to Economic Analysis, 172

1988 xvi + 416 pages  
ISBN 0-444-70429-6  
Forthcoming

This book arose out of research carried out by the authors in the period 1983–1987 whilst at the National Institute of Economic and Social Research. A number of things combined to impart the basic thrust of the research: partly the developments in formulating and estimating rational expectations models, and partly actual developments in the UK economy itself.

### Explaining the Growth of Government

Edited by J.A. Lybeck and M. Henrekson

Contributions to Economic Analysis, 170

1988 viii + 396 pages  
ISBN 0-444-70426-4  
Forthcoming

This book explains the post-war growth of the public sector in a number of developed economies. The purpose is to see whether scientists familiar with their respective countries' institutional, political and economic framework, but still working as a group, can advance some common factors behind the growth of government.

# North-Holland

In the U.S.A. and Canada:  
Elsevier Science Publishing Co. Inc.,  
P.O. Box 1663, Grand Central Station,  
New York, NY 10163, U.S.A.

In all other countries:  
Elsevier Science Publishers,  
Book Order Department,  
P.O. Box 211,  
1000 AE Amsterdam, The Netherlands.

US \$ prices are valid only in the USA and Canada. In all other countries the Dutch Guilder (Dfl.) price is definitive. Customers in the Netherlands, please add 6% B.T.W. In New York State applicable sales tax should be added. All prices are subject to change without prior notice.

NH/40q/BK/0971

# JOB OPENINGS FOR ECONOMISTS

Available only to AEA members and institutions that agree to list their openings.

## Annual Subscription Rates

U.S.A., Canada, and Mexico (first class): \$15.00, regular AEA members and institutions  
\$ 7.50, junior members of AEA  
All other countries (air mail): \$22.50, regular AEA members and institutions  
\$15.00, junior members of AEA

Please begin my issues with:

☐ February ☐ April ☐ June ☐ August ☐ October ☐ December

Name \_\_\_\_\_  
First Middle Last

Address \_\_\_\_\_

\_\_\_\_\_  
City State/Country Zip/Postal Code

Check one:

- ☐ I am a member of the American Economic Association.  
☐ I would like to become a member. My application and payment are enclosed.  
☐ (For institutions) We agree to list our vacancies in JOE.

Send payment (U.S. currency only) to:

THE AMERICAN ECONOMIC ASSOCIATION  
1313 21st Avenue South  
Nashville, Tennessee 37212

## Computer Access to Articles in the JEL Subject Index

Online computer access to the *JEL* and *Index of Economic Articles* database of journal articles is currently available through DIALOG Information Retrieval Service. DIALOG file 139 (*Economic Literature Index*) contains complete bibliographic citations to articles from the nearly 300 journals listed in the quarterly *JEL* issues from 1969 through the current issue. The abstracts published in *JEL* since June 1984 are also available as part of the full bibliographic record. The *Economic Literature Index* also includes citations to articles in the 1979 and 1980 collective volumes (collected papers, proceedings, etc.) for the *Index* database; other years will be added as soon as completed. The file may be searched using free-text searching techniques or author, journal, title, geographic area, date, and other descriptors, including descriptor codes based on the *Index's* four-digit subject classification numbers. (For a complete description of the *Economic Literature Index* with search examples and suggestions for searching techniques, see the article "Online Information Retrieval for Economists—The Economic Literature Index," in the December 1985 issue of the *Journal of Economic Literature*.)

### *Access Options:*

- **DIALOG** offers a variety of contract choices, including the option (for a low annual fee) to pay for only what you use. Most university libraries already subscribe to DIALOG. For information on the DIALOG service, contact your librarian or write to or call: DIALOG Information Services, Inc., Marketing Department, 3460 Hillview Avenue, Palo Alto, California 94304 (800-3-DIALOG or 800-334-2564).
- **Knowledge Index**, a DIALOG service available after 6 p.m. and on weekends, may be accessed at the low rate of \$24/hour, charged to a major credit card. A one time start-up fee of \$35.00 buys 2 hours free time during the first month after log-on. Call 800-3-DIALOG for information.
- **EasyNet**, a gateway service, provides menus to guide the untrained user through database searches in DIALOG and other databases. For information, call 1-800-841-9553 or dial up EasyNet on your terminal (1-800-EASYNET) and pay for your search by credit card.

### *Classroom Instruction:*

- DIALOG's Classroom Instruction Program, available at a special rate of \$15/connect hour to academic institutions for supervised instruction, permits teachers to incorporate online bibliographic searching in their courses. For information, contact DIALOG or your librarian.

# Journal of International Economic Integration

**Solicits Papers to Compete for the  
Annual Daeyang Prize in Economics (\$7,000)  
and Welcomes Subscriptions by Interested Parties**

## **Current issues include**

**Bela Balassa**, *Japanese Trade Policies Towards Developing Countries.*

**Basant K. Kapur**, *Open-Economy Response to a Terms of Trade Shock in a Growth Context.*

**Norman C. Miller**, *A General Approach to the Balance of Payments and Exchange Rates.*

**Leonard F.S. Wang**, *Product Market Imperfections and Customs Unions theory.*

**Gene M. Grossman**, *The Employment and Wage Effects of Import Competition in the United States.*

**Wilfred J. Ethier and Ronald D. Fischer**, *The New Protectionism*

**Joshua Alzenman**, *Inflation, Tariffs and Tax Enforcement Costs*

**Chung H. Lee and Selji Naya**, *The Internationalization of U.S. Service Industries and Its Implications for Developing Countries.*

The Journal of International Economic Integration is published biannually (Spring and Autumn) by the Institute for International Economics, King Sejong University, Seoul, Korea.

The purpose of the Journal is to support and encourage research in the area of international trade, international finance and other related economic issues that include general professional interest in international economic affairs. Welcoming both theoretical and empirical analyses in international economics, the Journal is strongly interested in the issues of the international economic cooperation.

- The Journal welcomes unsolicited manuscripts, which will be considered for publication by the Editorial Board.
- From papers selected for publication, the Prize committee will choose the best manuscript(s) to receive the \$7,000 Daeyang Prize. The winner of the prize is announced in the Spring issue every year.
- The manuscripts should be accompanied by an abstract of no more than 100 words and a brief curriculum vitae containing the author's academic career. All submissions should be typewritten, double-spaced, in English with footnotes, references, figures, tables and any other illustrative material on separate sheets.
- Three copies of the manuscript and all accompanying material should be submitted to the following address by October 31, 1988 for consideration for 1989 publication.
- For subscriptions to the Journal (\$20 per year for individuals, \$30 per year for institutions), send a check or money order payable to King Sejong University to the following address.

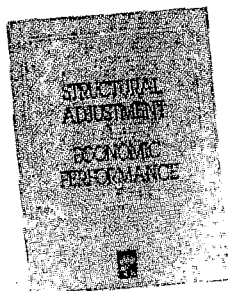
**Institute for International Economics  
King Sejong University  
Seongdong-Ku, Seoul, Korea**



O E C D



O C D E



### **Structural Adjustment and Economic Performance**

The potential for stronger economic growth in OECD countries is better than at any recent time. To exploit this fully requires adjustment to changing conditions and exploiting new opportunities. This report asks how this can be done. Concentrating on micro-economic policies, it examines the reasons for the outstanding growth of the 1950s and 1960s, and it sets out a program for policy reform in many areas. It provides a comprehensive review of a broad range of public policies in the advanced economies and analyses their economic consequences. The book is packed with statistical information, including over 100 statistical tables and over 40 graphs, some of which are presented in full color.

30-87-02-1, March 1988, 371 pages, ISBN 92-64-13006-3, \$39.95

### **National Accounts Volume 1: Main Aggregates 1960-1986.**

The 1988 edition of one of OECD's most asked-for publications. Contains graphs for each OECD country showing GDP, Private and Government Final Consumption Expenditure, and Gross Fixed Capital Formation; tables for each country showing the main aggregates in national currencies; "growth triangles" showing percent changes for the main comparative tables in U.S. dollars and in Purchasing Power Parities.

30-88-01-3, February 1988, 151 pages, ISBN 92-64-03017-4, \$27.00

### **External Debt Statistics.**

This report, containing statistics on the volume and composition of the external debt of 155 countries in 1985 and 1986, covers more countries than any other publication of its kind. The way in which the figures were compiled enables the reader to make more comparisons than is usually possible. The report also includes estimates of the amortisation payments each country was due to make on its long-term debt in 1987. Full technical explanations are provided.

43-87-05-1, January 1988, 29 pages, ISBN 92-64-13040-3, \$11.00

### **The Costs of Restricting Imports: The Automobile Industry.**

Presents the findings of studies by independent analysts of the effects of restrictions on imports and sales of foreign cars in four OECD countries: the United States, Canada, France, and the United Kingdom. It also demonstrates the usefulness of a checklist devised by the OECD and recommended to governments of Member countries in 1985 to help them assess the impact of proposed and existing regulations on trade in all products.

24-87-06-1, January 1988, 173 pages, ISBN 92-64-13037-3, \$18.00

To order, send your check or money order to:

### **OECD Publications and Information Center**

2001 L Street, NW, Washington, DC 20036-4095

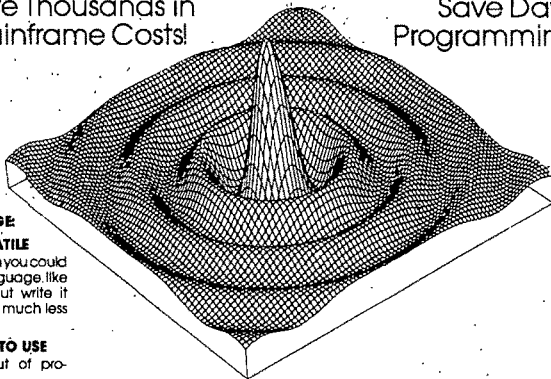
Telephone: (202) 785-6323

# GAUSS

## The New Standard for Scientific and Statistical Computation.

Save Thousands in  
Mainframe Costs!

Save Days in  
Programming Time!



### SOME FEATURES OF THE GAUSS PROGRAMMING LANGUAGE:

#### FULL-FEATURED AND VERSATILE

Write essentially any program you could write in a conventional language, like FORTRAN, PASCAL, or C, but write it faster, more easily, and with much less code.

#### EASY TO LEARN AND EASY TO USE

GAUSS takes the work out of programming.

#### EXTREMELY FAST

GAUSS provides the fastest computation, and the fastest I/O, of any program available for PC's, by far. On a plain PC, invert a 50x50 matrix in 14.66 seconds, compute the sums and means of variables in a data set with 10,000 observations, 50 variables in under 1 minute.

#### POWERFUL

GAUSS provides the basic tools of modern applied mathematics, and makes it easy to apply those tools.

#### NUMERICALLY ACCURATE

GAUSS uses state of the art numerical algorithms (LINPACK, EISPACK), and in addition takes optimal advantage of the extended precision of the 8087 numeric coprocessor.

#### COMMENTS FROM GAUSS USERS

"I used to use FORTRAN and PASCAL for languages, TSP and Minitab for statistics, MATLAB for math, and NAG and IMSL for FORTRAN subroutines. Now I just use GAUSS."  
Dr. Choon-Geol Moon  
Rutgers University

"GAUSS is the most beautifully designed software I have ever seen."  
Professor Warren Sanderson  
SUNY Stony Brook

## The GAUSS Mathematical & Statistical System

### • DATABASE MANAGEMENT (enter, convert, edit, sort, merge)

• **STATISTICS** (means, frequencies, crosstabs, regression, non-parametrics, general max likelihood, non-linear least squares, simultaneous equations, logit, probit, loglinear models, & more)

• **PUBLICATION QUALITY GRAPHICS** (2D & 3D; color, hidden line removal, zoom, pan; up to 4096 x 3120 resolution; produce Tektronix format files; output to most screen drivers, plotters, printers)

### • PLUS:

• SIMULATION • TIME SERIES/SIGNAL PROCESSING

• LINEAR PROGRAMMING • NON-LINEAR OPTIMIZATION

• NON-LINEAR EQUATION SOLUTION

• INTERACTIVE MATRIX PROGRAMMING

• LARGE-SCALE MODULAR PROGRAMMING

• ADD YOUR OWN COMMANDS

• LINK FORTRAN, C, ASSEMBLER SUBROUTINES

### OTHER FEATURES OF THE GAUSS PROGRAMMING LANGUAGE INCLUDE:

- full-screen editor, screen, printer, file, and keyboard I/O
- specialized functions for statistics and data handling
- state-of-the-art random number generators
- complex arithmetic; polynomial operators; trig functions
- probability density and cumulative distribution functions
- sequence functions, functions for recursive series
- arithmetic, relational, and logical operators
- numerical integration and differentiation
- UNPACK, EISPACK, and related algorithms, including LU, Cholesky, QR, and SVD decompositions; general and positive definite inverses; pseudo inverse; general and positive definite equation solutions; real general and symmetric eigenvalues and eigenvectors.
- Coded mostly in Assembler, core program smaller than 200K

Call or Write:

**APTECH  
SYSTEMS, INC.**

1914 N. 34th St., Suite 301  
Seattle, WA 98103  
(206) 547-1733

# THE GAUSS

## MATHEMATICAL AND STATISTICAL SYSTEM

for IBM PC-XT-AT-System/2 and Compatibles  
written by Lee E. Edlefsen and Samuel D. Jones

Buy the GAUSS Programming Language by itself or as part of the GAUSS Mathematical and Statistical System, which includes 2D & 3D graphics plus over 200 applications programs written in the GAUSS Programming Language for doing a variety of mathematical, statistical, and scientific tasks. Full source code is provided for these programs.

### 30 DAY MONEY-BACK GUARANTEE

The GAUSS Mathematical and Statistical System	\$350
The GAUSS Programming Language (alone)	\$200
Shipping/handling, continental USA, 2nd Day Air	\$8.50
Shipping/handling, continental USA, Ground	\$5.00
GAUSS requires 320K (512K required for high resolution graphics) DOS 2.10+, and a math coprocessor.	

NOT COPY PROTECTED

Please mention this publication when responding to this ad.

AR-12/87

# The American Economic Review

## PAPERS AND PROCEEDINGS

OF THE

**One-Hundredth Annual Meeting**

OF THE

AMERICAN ECONOMIC ASSOCIATION

**Chicago, Illinois, December 28–30, 1987**

**Program Arranged by Robert Eisner**

**Papers and Proceedings Edited by Harvey S. Rosen and Wilma St. John**

**MAY 1988**

# THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

## Officers

### *President*

ROBERT EISNER  
Northwestern University

### *President-elect*

JOSEPH A. PECHMAN  
The Brookings Institution

### *Vice-Presidents*

MARTIN S. FELDSTEIN  
National Bureau of Economic Research  
and Harvard University  
F. M. SCHERER  
Swarthmore College

### *Secretary-Treasurer*

C. ELTON HINSHAW  
Vanderbilt University

### *Editor of The American Economic Review*

ORLEY C. ASHENFELTER  
Princeton University

### *Editor of The Journal of Economic Literature*

JOHN PENCAVEL  
Stanford University

### *Editor of The Journal of Economic Perspectives*

JOSEPH E. STIGLITZ  
Princeton University

## Executive Committee

### *Elected Members of the Executive Committee*

SHERWIN ROSEN  
University of Chicago  
THOMAS J. SARGENT  
University of Minnesota  
ROBERT J. BARRO  
Harvard University  
JUDITH A. THORNTON  
University of Washington  
GEORGE A. AKERLOF  
University of California-Berkeley  
ISABEL V. SAWHILL  
Urban Institute

### *EX OFFICIO Members*

ALICE M. RIVLIN  
The Brookings Institution  
GARY S. BECKER  
University of Chicago

●Printed at Banta Company, Menasha, Wisconsin.

●Copyright © American Economic Association 1988. All rights reserved.

●No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, subscriptions, and changes of address, should be sent to the American Economic Association, 1313 21st Avenue South, Suite 809, Nashville, TN 37212-2786. Please remit membership payment with the application included elsewhere in this journal. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

*THE AMERICAN ECONOMIC REVIEW* (ISSN 0002-8282), May 1988, Vol. 78, No. 2, is published five times a year (March, May, June, September, December) by the American Economic Association, 1313 21st Avenue South, Suite 809, Nashville, TN 37212-2786. Annual subscription fees: Institutional subscriber-\$125.00, Individual subscriber-\$72.00, Regular members-\$38.50, \$46.20, or \$53.90 depending on income. A subscription also includes the *Journal of Economic Literature* and the *Journal of Economic Perspectives*. In countries other than the U.S.A., add \$16.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. POSTMASTER: Send address changes to the *American Economic Review*, 1313 21st Avenue South, Suite 809, Nashville, TN 37212-2786.

# THE AMERICAN ECONOMIC REVIEW

---

VOL. 78 NO. 2

MAY 1988

---

*PAPERS AND PROCEEDINGS*

OF THE

*One-Hundredth Annual Meeting*

OF THE

AMERICAN ECONOMIC ASSOCIATION

Chicago, Illinois

December 28–30, 1987

*Program Arranged by Robert Eisner*

*Papers and Proceedings Edited by Harvey S. Rosen and Wilma St. John*

Copyright © AMERICAN ECONOMIC ASSOCIATION, 1988



## CONTENTS

<b>Editors' Introduction</b> .....	<i>Harvey S. Rosen and Wilma St. John</i>	vii
------------------------------------	---	-----

## PAPERS

<b>Richard T. Ely Lecture</b>		
The Challenge of High Unemployment .....	<i>Alan S. Blinder</i>	1
<b>The Political Economy of Terrorism</b>		
To Bargain or Not To Bargain: That Is The Question .....	<i>Harvey E. Lapan and Todd Sandler</i>	16
Free Riding and Paid Riding in the Fight Against Terrorism .....	<i>Dwight R. Lee</i>	22
Intervention Policy Analysis of Skyjackings and Other Terrorist Incidents .....	<i>Jon Cauley and Eric Iksoon Im</i>	27
<b>The Natural Rate Theory Reconsidered</b>		
The Persistence of Unemployment .....	<i>Robert J. Barro</i>	32
Long-Term Unemployment and Macroeconomic Policy ..	<i>Assar Lindbeck and Dennis J. Snower</i>	38
Fairness and Unemployment .....	<i>George A. Akerlof and Janet L. Yellen</i>	44
<b>The Economics of the Arms Race</b>		
Self-Interest and National Security .....	<i>William G. Shepherd</i>	50
Economics Consequences of the Arms Race: The Second-Rate Economy .....	<i>Seymour Melman</i>	55
U.S. Military Power, the Terms of Trade, and the Profit Rate .....	<i>Tom Riddell</i>	60
<b>The Economics of the Aging of the Baby Boom</b>		
The Baby Boom's Legacy: Relative Wages in the Twenty-First Century .....	<i>Phillip B. Levine and Olivia S. Mitchell</i>	66
The Baby Boom, Housing, and Financial Flows .....	<i>Joyce Manchester</i>	70
Social Security Benefits and the Baby-Boom Generation .....	<i>Tabitha A. Doescher and John A. Turner</i>	76
<b>The Feminization of Poverty</b>		
Child Support Payments: Evidence from Repeated Cross Sections .....	<i>Andrea H. Beller and John W. Graham</i>	81
Getting into Poverty Without a Husband, and Getting Out, With or Without .....	<i>Thomas J. Kniesner, Marjorie B. McElroy, and Steven P. Wilcox</i>	86
Poverty Among Women and Children: What Accounts for the Change? .....	<i>Laurie J. Bassi</i>	91
<b>Markets for Information</b>		
Selling and Trading on Information in Financial Markets ...	<i>Anat R. Admati and Paul Pfleiderer</i>	96
Informational Theories of Employment .....	<i>Beth Allen and Costas Azariadis</i>	104
Parallel Search and Information Gathering .....	<i>Tara Vishwanath</i>	110
<b>Challenging Some Conventional Wisdoms about the Labor Market</b>		
The Un-Natural Rate of Unemployment: An Econometric Critique of the NAIRU Hypothesis .....	<i>David M. Gordon</i>	117
The Growth of Low-Wage Employment, 1963-86 .....	<i>Barry Bluestone and Bennett Harrison</i>	124
The Reemergence of Segmented Labor Market Theory .....	<i>William T. Dickens and Kevin Lang</i>	129

**Constitutional Convention Bicentennial**

- Contractarian Political Economy and Constitutional Interpretation . . . . . *James M. Buchanan* 135  
 Original Intent, History, and Doctrine: The Constitution and Economic Liberty . . . . . *Harry N. Scheiber* 140  
 Contested Exchange: Political Economy and Modern Economic Theory . . . . . *Samuel Bowles and Herbert Gintis* 145

**Issues in the Black Community**

- Income, Wealth, and Investment Behavior in the Black Community . . . . . *Andrew F. Brimmer* 151  
 The Social Preference for Fair Housing: During the Civil Rights Movement and Since . . . . . *Wilhelmina A. Leigh* 156

**Uncertainty in Macroeconomics**

- Uncertainty Across Models . . . . . *Christopher A. Sims* 163  
 The Fate of Systems with "Adaptive" Expectations . . . . . *Albert Marcet and Thomas J. Sargent* 168  
 Consumption: Beyond Certainty Equivalence . . . . . *Olivier Jean Blanchard and N. Gregory Mankiw* 173  
 Macroeconomic Implications of the Information Revolution . . . . . *George M. von Furstenberg and Esfandiar Maasoumi* 178

**Why is Unemployment So High in Europe?**

- Beyond the Natural Rate Hypothesis . . . . . *Olivier Jean Blanchard and Lawrence H. Summers* 182  
 Is European Unemployment Classical or Keynesian? . . . . . *Robert M. Coen and Bert G. Hickman* 188  
 West European Unemployment: Corporatism and Structural Change . . . . . *Andrew Glyn and Bob Rowthorn* 194

**Tax Policy and Investment: A Reconsideration**

- Investment, Financing Decisions, and Tax Policy . . . . . *Steven Fazzari, R. Glenn Hubbard, and Bruce Petersen* 200  
 Business Tax Policy, The Lucas Critique, and Lessons from the 1980's . . . . . *Robert S. Chirinko* 206  
 Investment Tax Incentives and Frequent Tax Reforms . . . . . *Alan J. Auerbach and James R. Hines, Jr.* 211

**Technological Innovation and Productivity Change in Japan and the United States**

- Productivity and Economic Growth in Japan and the United States . . . . . *Dale W. Jorgenson* 217  
 Industrial R&D in Japan and the United States: A Comparative Study . . . . . *Edwin Mansfield* 223  
 Why are Americans Such Poor Imitators? . . . . . *Nathan Rosenberg and W. Edward Steinmueller* 229

**Differing Theoretical Perspectives on the Household**

- Gender Difference: The Role of Endogenous Preferences and Collective Action . . . . . *Elaine McCrate* 235  
 Tied Transfers and Paternalistic Preferences . . . . . *Robert A. Pollak* 240  
 Risk, Private Information, and the Family . . . . . *Mark R. Rosenzweig* 245

**High School Economics: Implications for College Instruction**

- A Report Card on the Economic Literacy of U.S. High School Students . . . . . *William B. Walstad and John C. Soper* 251  
 Variables Affecting Success in Economic Education: Preliminary Findings from a New Data Base . . . . . *William J. Baumol and Robert J. Highsmith* 257  
 The Effects of Advanced Placement on College Introductory Economics Courses . . . . . *Stephen Buckles and John S. Morton* 263

**Inflation and Full Employment**

- Fluctuations in Equilibrium Unemployment . . . . . *Robert E. Hall* 269  
 The Role of Wages in the Inflation Process . . . . . *Robert J. Gordon* 276

**The Measure and Character of American Unemployment**

The Measurement of Unemployment . . . . .	<i>Janet L. Norwood</i>	284
What Is So Natural About High Unemployment? . . . . .	<i>Richard S. Krashevski</i>	289
Evaluating the European View that the United States Has No Unemployment Problem . . . . .	<i>Richard B. Freeman</i>	294

**Economic Aspects of Product Liability**

Product Liability and Regulation: Establishing the Appropriate Institutional Division of Labor . . . . .	<i>W. Kip Viscusi</i>	300
The Political Economy of Workers' Compensation: Lessons for Product Liability . . . . .	<i>Patricia M. Danzon</i>	305
The Political Economy of Product Liability Reform . . . . .	<i>Richard A. Epstein</i>	311

**Surprises from Deregulation**

Surprises of Airline Deregulation . . . . .	<i>Alfred E. Kahn</i>	316
Surprises from Telephone Deregulation and the AT&T Divestiture . . . . .	<i>Robert W. Crandell</i>	323
Interaction of Financial and Regulatory Innovation . . . . .	<i>Edward J. Kane</i>	328

**The Widespread Depression Overseas: American and Pacific Influences**

On Macroeconomic Implications of Price Setting in the Open Economy . . . . .	<i>Jean-Paul Fitoussi and Jacques Le Cacheux</i>	335
Exchange Rates, Wages, and the International Allocation of Capital . . . . .	<i>Slobodan Djajić</i>	341
A Working Model of Slump and Recovery from Disturbances to Capital-Goods Demand in an Open Nonmonetary Economy . . . . .	<i>Edmund S. Phelps</i>	346

**Search Behavior in Labor and Product Markets**

Pareto Inefficiency of Market Economies: Search and Efficiency Wage Models . . . . .	<i>Bruce Greenwald and Joseph E. Stiglitz</i>	351
"High-Low Search" in Product and Labor Markets . . . . .	<i>Steve Alpern and Dennis J. Snower</i>	356
The Search Equilibrium Approach to Fluctuations in Employment . . . . .	<i>Christopher A. Pissarides</i>	363
Self-Fulfilling Optimism in a Trade-Friction Model of the Business Cycle . . . . .	<i>Allan Drazen</i>	369

**"The New Industrial State" After Twenty Years**

Time and the New Industrial State . . . . .	<i>John Kenneth Galbraith</i>	373
Discussion . . . . .	<i>Barry Bluestone</i>	377
	<i>Robert M. Solow</i>	378
	<i>F. M. Scherer</i>	380

**Efficiency Wages, Labor Relations, and Full Employment**

Relative Wages, Efficiency Wages, and Keynesian Unemployment . . . . .	<i>Lawrence H. Summers</i>	383
Unemployment, Labor Relations, and Unit Labor Costs . . . . .	<i>James B. Rebitzer</i>	389
Labor Discipline and Aggregate Demand: A Macroeconomic Model . . . . .	<i>Samuel Bowles and Robert Boyer</i>	395

**What Do We Know About Consumption and Saving, and What are the Implications for Fiscal Policy?**

Consumption, Saving, and Fiscal Policy . . . . .	<i>Michael J. Boskin</i>	401
Consumption, Computation Mistakes, and Fiscal Policy . . . . .	<i>Laurence J. Kotlikoff, William Samuelson, and Stephen Johnson</i>	408
Are Consumers Forward Looking? Evidence from Fiscal Experiments . . . . .	<i>James M. Poterba</i>	413

**The Estimation and Measurement of Spillover Effects of R&D Investment**

Industry Effects and Appropriability Measures in the Stock Market's Valuation of R&D and Patents . . . . .	<i>Iain Cockburn and Zvi Griliches</i>	419
Appropriability, R&D Spending, and Technological Performance . . . . .	<i>Richard C. Levin</i>	424
Interindustry R&D Spillovers, Rates of Return, and Production in High-Tech Industries . . . . .	<i>Jeffrey I. Bernstein and M. Ishaq Nadiri</i>	429



**Is It Money or Credit, or Both, or Neither?**

Credit, Money, and Aggregate Demand . . . . .	<i>Ben S. Bernanke and Alan S. Blinder</i>	435
Monetary Policy Without Quantity Variables . . . . .	<i>Benjamin M. Friedman</i>	440
Money and Credit in the Monetary Transmission Process . .	<i>Karl Brunner and Allan H. Meltzer</i>	446

**Comparative Strategies for Economic Reform**

Economic Reforms Within and Beyond the State Sector . . . . .	<i>Tamás Bauer</i>	452
On the Strategy for Implementing Economic Reform in the USSR . . . . .	<i>Valery L. Makarov</i>	457
Choosing a Strategy for China's Economic Reform . . . . .	<i>Jinglian Wu and Bruce L. Reynolds</i>	461

**Applications of International Comparisons of Prices and Quantities**

What We Have Learned about Prices and Quantities from International Comparisons: 1987 . . . . .	<i>Alan Heston and Robert Summers</i>	467
National Price Levels and the Prices of Tradables and Nontradables . . . . .	<i>Irving B. Kravis and Robert E. Lipsey</i>	474
The Sensitivity of International Comparisons of Capital Stock Measures to Different "Real" Exchange Rates . . . . .	<i>Edward E. Leamer</i>	479

**PROCEEDINGS**

John Bates Clark Award . . . . .	486
Minutes of the Annual Meeting . . . . .	487
Minutes of the Executive Committee Meetings . . . . .	488

**Reports**

Secretary . . . . .	<i>C. Elton Hinshaw</i>	495
Treasurer . . . . .	<i>Rendigs Fels</i>	499
Finance Committee . . . . .	<i>Rendigs Fels</i>	501
Editor, <i>American Economic Review</i> . . . . .	<i>Orley Ashenfelter</i>	502
Editor, <i>Journal of Economic Literature</i> . . . . .	<i>John Pencavel</i>	511
Editor, <i>Journal of Economic Perspectives</i> . . . . .	<i>Joseph Stiglitz</i>	513
Director, <i>Job Openings for Economists</i> . . . . .	<i>C. Elton Hinshaw</i>	516
Committee on Economic Education . . . . .	<i>W. Lee Hansen</i>	518
Committee on the Status of Women in the Economics Profession . . . . .	<i>Nancy M. Gordon</i>	520
Committee on U.S.-Soviet Exchanges . . . . .	<i>Franklyn D. Holzman</i>	522
Committee on U.S.-China Exchanges in Economics . . . . .	<i>Gregory C. Chow</i>	523
Representative to the National Bureau of Economic Research . . . . .	<i>David Kendrick</i>	525
Representative to the American Association for the Advancement of Science . . . . .	<i>Adam Rose</i>	527

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

## Editors' Introduction

This volume contains the *Papers and Proceedings* of the one-hundredth annual meeting of the American Economic Association. The *Proceedings* record the business activities of the Association in 1987; the annual membership meeting; the March and December meetings of the Association's officers and committees. The *Papers* constitute the greater part of the volume. They comprise eighty-four contributions that fill roughly the same number of pages as two regular issues of the *American Economic Review*. We would like to take this opportunity to answer a number of commonly asked questions about the *Papers*.

**Who chooses the authors?** About a year in advance, the Association's President-elect, acting as program chairman, decides on the topics for which sessions will be organized. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. (A *Call for Papers* was published annually in the Notes section of the *AER*, and now appears in the Fall issue of the *Journal of Economic Perspectives*.) The President-elect invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the theme of the session, and asks others to give comments on the papers. The program chairman decides at the time of organization which sessions are to be included in this volume. Space limitations restrict the number of printed sessions. This year we are printing 29 sessions, although a total of 142 sessions were sponsored, either solely by the American Economic Association or jointly with other allied societies.

**Are discussants' comments published?** There has been no standard practice with regard to the publication of comments and discussions in the past. This year the President-elect decided to publish one set of com-

ments for one panel session. For the other sessions, names and affiliations of commentators are printed at the start of each session, permitting readers especially interested in particular comments to write to the commentator for a copy of the discussion.

**What standards must the papers meet?** The guidelines under which papers are published in the *Papers and Proceedings* differ considerably from those governing regular issues of the *Review*. First, the length of papers is strictly controlled. Except in unusual circumstances they must be no more than twelve typescript pages in three-paper sessions, and eighteen typescript pages in two-paper sessions. Second, papers are not subjected to a formal refereeing process. However, a paper can be rejected if, after reading it, we conclude that it is utterly without merit. This year we are pleased to report that no paper has been rejected on this ground. Third, their content and range of subject matter reflect the wishes of the President-elect to investigate and expose the current state of economic research and thinking. In most cases they are therefore exploratory and discursive, rather than formal presentations of original research.

In order to produce this volume by May, very rigid deadlines must be met and there is no time for communication with every author about editing changes made in order to improve content and style, and to satisfy space restrictions. Every effort is made to notify an author prior to the deadline if the paper is too long, or does not satisfy other specifications.

This year, most authors cooperated very nicely. We thank them for making our lives easier.

HARVEY S. ROSEN  
WILMA ST. JOHN

## RICHARD T. ELY LECTURE

# The Challenge of High Unemployment

By ALAN S. BLINDER\*

The Ely Lecture is an occasion to indulge in big think, to eschew equations and "speak prose"—a respite from the daily grind of vector autoregressions, Euler equations, and phase diagrams. I intend to take full advantage of this privilege tonight. Judging by past Ely Lectures, it is also an occasion either to celebrate the profession (or one's own contributions to it) or to chide it. Some combination of flaws in my character and flaws in our discipline incline me toward the latter.

My topic is the challenge of high unemployment, one which both policymakers and economists have failed to meet. The challenge to policymakers is to reduce unemployment. About this, I will be brief and to the point. The challenge to economists is to explain high unemployment and understand its implications for things economic. Here I will dwell longer.

### I. The Challenge to Policymakers

The failure to provide productive employment for all those willing and able to work has long been one of the major weaknesses of market capitalism. Since the mid-1970's, it has been shamefully debilitating. If one picture is worth a thousand words, Figure 1 will help shorten the lecture. It shows unemployment rates in the United States and the European OECD countries in two different periods: 1961–74 and 1975–85. The contrast is stark.

The costs summarized in this graph are enormous for the United States and colossal



FIGURE 1

for Europe. And the corresponding Okun gaps, wide as they are, understate the full costs. A high-pressure economy provides opportunities, facilitates structural change, encourages inventiveness and innovation, opens doors for society's underdogs, and yields a fiscal dividend that can be spent, among other things, on public charity. All these promote the social cohesion and economic progress that make democratic mixed capitalism such a wonderful system when it works well. A low-pressure economy slams the doors shut, breeds a bunker mentality that resists change, stifles productivity growth, and fosters both inequality and mean-spirited public policy. All this makes reducing high unemployment a political, economic, and moral challenge of the highest order.

To make the point in extreme form, think about the U.S. economy during World War II, when unemployment virtually vanished, the economy flexed its muscles, and America truly became a land of opportunity. Among the remarkable features of this period was a 16 percent rise in real consumer spending between 1939 and 1944 despite the wholesale redirection of economic activity toward war

\*Department of Economics, Princeton University, Princeton, NJ 08544. I am grateful to Will Baumol, Ben Bernanke, Avinash Dixit, Bob Eisner, Steve Goldfeld, Bob Gordon, Dan Hamermesh, David Romer, Harvey Rosen, Bob Solow, and Larry Summers for helpful comments.

production. Now imagine that there was no war and all those soldiers and equipment went abroad to work, not to fight, sending home no goods, just remittances. But leave in your minds all the rationing and other nasty Harberger triangles caused by the shortage economy. Ask yourself whether the utility of the representative American would have been higher under these hypothetical conditions or under the actual conditions of 1939—or, abstracting from secular growth, even 1987 for that matter. My suggested answer, you can tell, is yes.

A debater's point, you say, for wartime unemployment rates were absurdly and unsustainably low. Probably so. But remember that just fourteen years ago the unemployment rates (using U.S. concepts) were 3.2 percent in the United Kingdom, 2.7 percent in France, and 0.7 percent in Germany. These are surely not unimaginable worlds. And think of the social dividend that would be reaped if those countries got unemployment even halfway back to where it was in 1973. Or think about the present-day United States. While many people see today's 6 percent national unemployment rate as "full employment," the unemployment rate is more like 3–4 percent in Massachusetts and New Jersey. Those two states and parts of others do show clear signs of labor scarcity: Help Wanted signs are everywhere and wages are rising faster than the national average. For all I know, there may even be people whose marginal utility of leisure exceeds their wage. But there are no signs of chaos, and shortages of goods and services are rare. The local economies are, as a matter of fact, doing quite well, thank you. Wouldn't it be nice if the whole country were in such good shape? Aren't we wasting something precious if it could be?

Yet in the United States and, especially, in Europe, those in authority often accept high unemployment with an air of resignation, as if it stemmed from acts of nature rather than from acts of man. This is an attitude conducive to paralysis; and so we wind up with an excess supply of excess supply.

The European and American experiences differ both quantitatively and qualitatively. While there is much we do not know about

the details, the broad outlines of the origins of high European unemployment are familiar enough. Intransigent trade unions and well-intentioned but unintelligent governments have erected a web of microeconomic barriers to full employment that both make labor more expensive and transform wages from variable into fixed costs. These include (with different weights in different countries) high minimum wages, excessive severance pay, heavy fixed costs of employment, restrictions on hiring and firing, support for the closed union shop, meaningless licensing requirements, heavy-handed workplace rules, and impediments to geographic mobility.<sup>1</sup> There is nothing at all "natural" about unemployment that results from such misguided micro policies, and economists rightly oppose them.

But there is also an important macro component to the slack we see in Europe today. And in the United States, which has avoided the horror stories of European labor markets, restrictive policy is virtually the whole story behind the Great Recession of the 1980's. Put plainly, governments here and abroad have used high unemployment to exorcise the inflationary demon.

Unfortunately, economists are terribly divided on the relative importance of the micro and macro explanations for high unemployment. Some think macro policy played a major role in the drama; others assign it only a bit part. This internal schism, I am afraid, contributes to the policy paralysis—which brings me to the role of economists, beginning with macroeconomists.

## II. The Challenge to Macroeconomists

Every science has its game playing and puzzle solving. It's harmless, good clean fun, helps sharpen the mind, and occasionally turns up something spectacularly useful. Economics is no exception, nor should it be. But I want to suggest that contemporary academic economists have taken a good thing

<sup>1</sup>Among the many possible references that could be cited, see the special 1986 supplement of *Economica* or John Gennard (1985).

too far, pushed the game-playing aspects beyond the region of even positive marginal returns, and disengaged themselves from the practical policy concerns that affect the lives of millions. We will not contribute much toward alleviating unemployment while we fiddle around with theories of Pareto optimal recessions—an avocation that might be called Nero-Classical Economics.

It wasn't always that way. A century ago, Alfred Marshall concluded the inaugural lecture for his chair at Cambridge with these words:

It will be my most cherished ambition...to increase the numbers of those whom Cambridge, the great mother of strong men, sends out into the world with cool heads but warm hearts, willing to give some at least of their best powers to grappling with the social suffering around them.<sup>2</sup>

Even after translating the soppy Victorian prose into the modern vernacular, Marshall's sentiments are frightfully out of touch with the realities of contemporary academia, where a stubborn fixation on the real world is apt to be considered boorish, if not downright anti-intellectual.

Yet is Marshall's ideal really foolishly romantic? Isn't it better than Nero's? Didn't Keynes have a point when he longed for the day when economists would be as useful as dentists? Greater concentration on real, rather than imagined, problems need not make economics less scientific. Why, for example, are so many scientists now working on AIDS and cancer? Yes, I know that part of the answer is the one Willy Sutton gave when asked why he robbed banks: "That's where the money is." But another part of the answer is: "That's where the suffering is." It's a good answer, too.

Don't get me wrong. I am not suggesting that we all forsake mathematics for social work. Being a do-gooder may not be the best

way to do good; nor should that be the sole concern of a scientist. Nor am I suggesting more top-notch, policy-oriented research will banish the scourge of high unemployment. Vested interests, ideological cant, and sheer ignorance surely hold more sway over policy than does economic science. I am suggesting something far more modest: that a major redirection of the work of hundreds of economists might conceivably raise the quality of national economic policy from, say, 3 to 4 on a scale of 10. Hell, Keynes did more than that by himself.

As I see it, the challenge of unemployment to macroeconomists is fourfold: to define involuntary unemployment, to explain it theoretically, to give the theory empirical content, and then to devise policies to reduce it.

#### A. First Challenge: Define It

Some economists, you know, lean toward the tautological view that anything done without literal compulsion must of necessity be voluntary. Others detect elements of involuntarism whenever constraints become too constraining. It may be that *involuntary unemployment* is like pornography: It's hard to define, but you know it when you see it.

Actually, defining involuntary unemployment is no trick at all in the mythical case of homogeneous labor. If labor supplied exceeds labor demanded at the going wage, the difference is literally and unambiguously involuntary. This simplistic view of the world identifies involuntary unemployment with wages that will not fall—a point to which I will return. But with heterogeneous labor the simple definition no longer works, and the whole concept gets slippery. What wage do we mean? Which types of labor?

In the Keynesian oral tradition, the term "involuntary unemployment" signifies two major ideas. The first is that there are identifiably bad times, called recessions or depressions, when the unemployment rate rises and signs of economic distress are apparent. The second, and more controversial, is that unemployment tends to be too high on average. Pursuing the analogy to pornography, perhaps we should treat the term involuntary unemployment as synonymous

<sup>2</sup>A. Marshall, "The Present Position of Economics," in Pigou (1925, p. 174), original in 1885. I thank Avinash Dixit for finding this quotation.

with "pornographic unemployment": joblessness without redeeming social value.

This suggests an operational definition. Ask the following simple question of job losers and job leavers: Would you willingly take your previous job back on the terms now available in the market? If the answer is yes, the person is involuntarily (or pornographically) unemployed. This seems a straightforward test whenever there is a well-defined previous job, but it cannot be readily applied to new entrants or reentrants.<sup>3</sup> Fortunately, job losers and job leavers constitute 60–70 percent of total measured unemployment in the United States and about 75–80 percent of the rise in unemployment during recessions.<sup>4</sup> So conceptual difficulties with new entrants and reentrants are of minor practical importance. We can probably get an excellent indication of *changes* in involuntary unemployment by looking only at job losers.

The definition helps distinguish involuntary (or socially useless) unemployment from voluntary (or socially useful) unemployment. People who are enjoying leisure rather than working at what they perceive as unusually low wages would not be considered involuntarily unemployed since they presumably would not take their old jobs back on the previous terms. But few of the unemployed seem to be doing that, and the facts that real wages are (a) close to a random walk<sup>5</sup> and (b) not very cyclical<sup>6</sup> cast serious doubt on the empirical importance of intertemporal substitution in labor supply. Similarly, people who are actively pursuing productive job

search are not uselessly unemployed. Certainly, there are such people; but probably not many. We know, for example, that the average job seeker spends only a few hours a week on search and rarely rejects a job offer.<sup>7</sup>

The mention of search brings up the second challenge: explaining high unemployment theoretically.

### B: Second Challenge: *Explain It Theoretically*

In my view, one main reason for our lack of progress in explaining high unemployment is that academic economists have spent too much time and energy debating whether involuntary (or pornographic) unemployment exists and too little theorizing why. Furthermore, too much of our theoretical debate has taken place within the confining strictures of homogeneous labor, where the question reduces to whether and why "the wage rate" is sticky. That is a reasonable question; but it is not the *only* question.

Once we force ourselves to think seriously about the *heterogeneity* of labor, the very concept of wage rigidity loses precision. For example, is it the *average level* of wages or the structure of *relative* wages that is sticky? Instead of sterile debates about why rational people would leave unexploited Harberger triangles lying on the table, we start thinking about things like relative status and coordination failures. These are important issues. I suspect they may be central to understanding high unemployment. But they simply cannot arise in a homogeneous labor market.

Let me illustrate by pursuing the tantalizing question raised by search theory: Why doesn't an unemployed person take the first job she finds while continuing to look for a better one? As a stylized example, why don't unemployed steelworkers go to work at McDonald's? And, if they do not, should we consider their unemployment voluntary?

The traditional search-theoretic answer is straightforward and almost certainly wrong.

<sup>3</sup>Kim Clark and Lawrence Summers (1979) have argued persuasively that many reentrants are really job losers. The definition also applies to such people.

<sup>4</sup>Data on unemployment by reason are available only for the last four recessions. In those recessions, job losers and leavers accounted for 70, 73, 93, and 80 percent of the peak-to-trough rise in the unemployment rate (using NBER cycle dates). The vast majority of this was from job losers. A regression of the job loser rate on time, a constant, and the overall unemployment rate (monthly data, January 1967 to February 1987) produces a coefficient of 0.75 on the latter.

<sup>5</sup>See Joseph Altonji and Orley Ashenfelter (1980).

<sup>6</sup>See Patrick Geary and John Kennan (1982); Mark Bils (1985).

<sup>7</sup>See Clark and Summers (pp. 54–55). Only 10 percent of unemployed people in the special 1976 job-search survey reported rejecting a job offer.

It holds that search is so much more efficient off the job than on the job that the efficiency gains from searching while unemployed outweigh the lost income. No evidence supports this hypothesis. We know that people can search better on the job in some labor markets. Even in markets where search is best done while unemployed, it is hard to believe that a few hours of search activity per week interfere unduly with holding a job—unless geographical relocation is necessary.

An alternative explanation posits the existence of substantial transactions costs from taking and leaving an interim job. On this view, the dislodged steelworker rationally refuses the job at McDonald's because his in-and-out costs exceed the value of the wages he could earn during a few weeks spent flipping hamburgers. This explanation is logically coherent and even believable for people who anticipate an extremely short spell of unemployment.<sup>8</sup> But most unemployment is accounted for by long spells. For example, 54 percent of all unemployment in 1984 was accounted for by those unemployed for 27 weeks or more.<sup>9</sup> And besides, it is hard to see how the in-and-out costs of taking a short-term job could possibly amount to much more than one day's time. That can hardly explain voluntarily forsaking several weeks' wages.

Another possibility is that workers who lose "good" jobs worry about being stigmatized by taking "bad" jobs. I could make this explanation sound less like pop sociology and more like modern economics by gussying it up with words like signalling, asymmetric information, and adverse selection. I could even say it with algebra—but not right after dinner. In whatever guise, the idea is simply that unemployed steelworkers do not want potential employers thinking of them as hamburger flippers. To those willing to venture beyond the confines of neoclassical economics, this is an appealing notion.

But there is one big problem. An unemployed steelworker can lose the stigma and keep the income by taking the McDonald's job, omitting it from his resume, and telling prospective employers that he is unemployed.

So let me suggest an alternative hypothesis based on a very old idea, one which all social scientists but economists find compelling and for which Robert Frank (1985), in particular, has argued eloquently: that people care deeply about their relative status in society. To be more precise, suppose utility depends not just on the *level* of income but also on one's *position in the income distribution*. Suppose further, and this is the critical leap, that you retain the relative status attached to your old job until you take a new one. Thus an unemployed steelworker remains a steelworker—both in his mind and in the minds of others—until he takes a new job; then his status changes. If concern about status is high enough and the gap between the available wage and unemployment compensation is low enough, the individual may prefer unemployment as a steelworker to employment on an inferior job.

Direct empirical evidence on this hypothesis is difficult to come by, though Frank has offered evidence for the importance of relative status in a wide variety of contexts, some of them even biological.<sup>10</sup> So, once again, a thought experiment may help. Suppose a plant closing costs a steelworker his job. After two weeks of puttering around the house, he walks past the local McDonald's and sees a Help Wanted sign. Does he walk in and take the job? I think not. Now *why* not? Is it because it would interfere with his job search? Not likely. Is it because he doesn't want personnel directors at other steel mills to think of him as a fast-food worker? Perhaps. But how would they know? I suggest that it may really be because he doesn't want his friends and neighbors—and,

<sup>8</sup>In such cases, intertemporal substitution is also an attractive explanation.

<sup>9</sup>See Summers (1986, Table 5, p. 353).

<sup>10</sup>Interviews conducted by Jean Baldwin Grossman (1980) found that most firms adjusted above-minimum wages promptly after the statutory increase in the minimum wage in January 1979.



especially; doesn't want *himself*—to see him in that low-status position.<sup>11</sup>

Though based on concern for social status rather on coordination failures, this idea is reminiscent of an old Keynesian saw: that workers resist wage reductions because they are concerned that other wages will not follow suit. To hone and quantify our intuition, consider the following simple example that applies to either case.

Utility for individual  $i$  depends on his own real income and on the ratio of his own wage to some comparison wage,  $w_i/w_j$ . Using Cobb-Douglas utility for convenience, utility while employed is

$$U = (w_i/w_j)^\alpha w_i^{1-\alpha} \equiv U_0.$$

Now suppose the worker loses his job and must choose between accepting a job paying  $\lambda w_i$  ( $\lambda < 1$ ) or remaining unemployed and receiving income  $b w_i$  ( $b < \lambda$ ) from unemployment insurance, home production, or whatever. If he takes the job, utility is

$$(1) \quad U = \left( \frac{\lambda w_i}{w_j} \right)^\alpha (\lambda w_i)^{1-\alpha} = \lambda U_0.$$

If he turns it down, he gets

$$(2) \quad U = \left( \frac{w_i}{w_j} \right)^\alpha (b w_i)^{1-\alpha} = b^{1-\alpha} U_0.$$

Thus he will prefer unemployment to the low-paying job if and only if

$$(3) \quad b^{1-\alpha} > \lambda.$$

When there is no concern for relative status, ( $\alpha = 0$ ), only income matters and the bad job is preferred to unemployment as long as  $b < \lambda$ . But as  $\alpha$  gets bigger, the left-hand side of (3) gets larger and the possibility that the worker might refuse the

TABLE 1—REPLACEMENT RATE NEEDED TO TURN DOWN JOB

	$\alpha = 0$	$\alpha = .2$	$\alpha = .5$	$\alpha = .8$
$\lambda = .90$	.90	.88	.81	.59
$\lambda = .80$	.80	.76	.64	.33
$\lambda = .70$	.70	.64	.49	.17
$\lambda = .50$	.50	.42	.25	.03
$\lambda = .30$	.30	.22	.09	.002

job grows. A convenient way to look at this is to ask how large  $b$  (the replacement rate) must be to induce the worker to turn down a job that offers a wage of  $\lambda w_i$ . Table 1 tabulates the answer for various combinations of  $\lambda$  and  $\alpha$ . For example, if  $\alpha = 0.2$ , the worker will turn down a job paying half his previous wage if his replacement rate is above 42 percent. The gap between 50 and 42 percent may not be exciting. But if  $\alpha$  is as large as  $1/2$ , the critical replacement rate drops to 25 percent—meaning that the worker prefers unemployment and a 75 percent drop in income to a job paying half his previous wage.

Precisely the same comparison arises in the Keynesian case of uncoordinated wage cutting. If workers assume that those earning  $w_j$  will not take a wage cut, they expect to receive (1) if they accept a  $100(1 - \lambda)\%$  wage cut, and (2) if they refuse and lose their jobs. Condition (3) is thus the condition for preferring a layoff to a wage cut when you do not expect other wages to fall. It turns out also to be the condition for refusing the wage cut when you *do* expect other wages to fall, for if you take a cut and retain your job, you get  $\lambda^{1-\alpha} U_0$ , while if you refuse and lose your job, you get  $(w_i/\lambda w_j)^\alpha (b w_i)^{1-\alpha} = b^{1-\alpha} \lambda^{-\alpha} U_0$ . The latter exceeds the former if and only if (3) holds.

Perusing Table 1 makes it clear that the value of  $\alpha$  is of great moment. If  $\alpha$  is small, concern for social status cannot take us very far in explaining unemployment. If  $\alpha$  is large, it becomes a powerful explanator. To "estimate"  $\alpha$ , I again ask you to introspect. Imagine that in one case your university raises *only your salary* by 10 percent, while in another it gives  $k$  percent *to everyone*. How large must  $k$  be for these two events to give

<sup>11</sup>As evidence for this, labor economists have found that high previous wages lead to high reservation wages. See, for example, Nicholas Kiefer and George Neumann (1979).

TABLE 2—UTILITY-EQUIVALENT RAISES

$\alpha = 0.2$		$\alpha = 0.5$		$\alpha = 0.8$	
(1)	(2)	(1)	(2)	(1)	(2)
5	6.3	5	10.3	5	27.6
10	12.7	10	21	10	61.1
15	19.1	15	32.3	15	101
20	25.6	20	44	20	149

Note: Shown in percent. Cols. (1) denote "Just for you," and Cols. (2) denote "For everyone."

you the same satisfaction? Table 2 shows some answers for several values of  $\alpha$  and raises of different sizes. For example, if  $\alpha = 0.2$ , a 10 percent raise given just to you is as good as a 12.7 percent raise across the board. Each of you can make your own judgment, but this strikes me as less concern with relative status than most real people have. Similarly, the  $\alpha = 0.8$  column strikes me as much more. Personal introspection tells me that  $\alpha$  is between 0.2 and 0.5. For example, if  $\alpha = 1/3$ , a 10 percent raise for me only makes me just as happy as a 15.4 percent raise for everyone in my university. That strikes me as roughly correct.

This much concern with relative status is enough to matter. For example, the entry that would appear in Table 1 for  $\alpha = 1/3$  and  $\alpha = 0.5$  is 0.35, meaning that I would rather accept unemployment and a 65 percent drop in income than take a job at half my present wage.

Now what I have just presented is an idea, not a model. It has been said that an economist is someone who sees that something works in practice and wonders if it also works in theory. I will not be so obtuse as to try to build a theoretical model incorporating this idea at this late hour. But the dim outlines of such a model are already implicit in an important recent paper by Laurence Ball and David Romer (1987b). Working with prices of goods rather than wages of labor, they show that a large real rigidity coupled with a small fixed cost of changing nominal prices can explain large nonneutralities of money. By analogy, I conjecture that it is possible to show that monetary shocks have large effects on employ-

ment when workers care about relative wages and firms have small fixed costs of changing nominal wages.

This is just one example of the possibilities that arise once we leave the mythical world of homogeneous labor—as I think we should. Happily, the latest developments in the never-ending quest for microfoundations of macroeconomics make heterogeneity an essential part of the story. I refer, in particular, to theories of unemployment based on imperfect information, efficiency wages, insider-outsider distinctions, and monopolistic competition. And I would like to see concern with relative wages and "fairness" included on this list, maybe at the top.

Models of labor markets with imperfect information stress such things as unobservable differences in productivity and inability of management to monitor the performance of individual workers. The central message of this burgeoning literature is that wages may not be able to clear markets because they are too busy doing other things. Insider-outsider models recognize the inherent asymmetry in the positions of incumbent workers and challengers. Heterogeneity of goods is, of course, the essence of monopolistic competition models. And efficiency-wage models provide many reasons why firms might deliberately set wages above market-clearing levels—for example, to reduce turnover or to encourage greater work effort.

Each of these approaches contributes something to giving theoretical coherence to the Keynesian intuition that unemployment is often too high. However, I do not wish to oversell the results, for the welfare economics is a bit dicey. In imperfect-information and efficiency-wage models, "too high" generally means higher than in some unattainable perfect-information equilibrium. In monopolistic competition models, output is lower than it would be under perfect competition. In these cases, policy interventions are not always called for and, if they are, may not take the form of macro stabilization policy. (See Ball and Romer, 1987a.) Still, I find all this a refreshing departure from the scholastic dogma of High Neoclassicism.

However, these new models have so far contributed little to explaining the *changes*

in unemployment that we observe in time-series and that we call business cycles. Indeed, some seem ill-suited to the task. Hysteresis models may be the most promising in this regard, especially in the European context, for they show how changes in demand can essentially drag supply along—in a neat reversal of Say's Law.

Finally, these models shed little light on why nominal shocks have strong real effects, for each is fundamentally a story about relative prices or real wages. As I just indicated, one way to transform a real rigidity into a nominal rigidity is to add costs of changing nominal prices or wages. George Akerlof and Janet Yellen (1985) add fixed costs of changing prices to a model with efficiency wages.<sup>12</sup> Olivier Blanchard and Nobuhiro Kiyotaki (1987), building on the insights of N. Gregory Mankiw (1985), do the same in a monopolistic competition model. I believe combining costs of changing money wages with a strong concern about relative wages and/or "fairness" is a promising approach to explaining how fluctuations in demand produce fluctuations in employment.

This theoretical work is still in its infancy (some of it is still *in utero*) and is not without difficulties. While costs of changing prices certainly exist, it is hard to believe that they are large. That is why Ball and Romer's (1987b) demonstration that large monetary nonneutralities can result from the interaction of small nominal rigidities and large real rigidities is so important. However, costs of changing *quantities* also undoubtedly exist; so it is not clear that adjustment costs logically lead to rigid prices and flexible quantities. Finally, theories based on fixed costs of changing prices ("menu costs") need to be recast in a dynamic framework which recognizes that optimal strategies are likely to be variants of the  $(S, s)$  rule of inventory theory in which firms adjust prices at different times.<sup>13</sup>

The Keynesian promised land is not yet in sight; but we may, at long last, be emerging from the arid desert and looking over the Jordan. Let me use the license granted me on this occasion to peer beyond where we can really see and speculate briefly on the outlines of a model that is both theoretically respectable and can be explained in mixed company without embarrassment. The model I envision—but do not have—has three main ingredients.

The first is efficiency wages, so there is no tendency for labor markets to clear in the naive neoclassical sense. Large firms, most of which have market power and some fat in their cost structures, pay wages high enough to maintain a queue of qualified job seekers and to retain the workers they have. They do so because turnover is disruptive, because higher wages attract superior applicants, and, perhaps most importantly, because workers perform better when they feel they are well paid.<sup>14</sup> The result is excess supply and unemployment in equilibrium. I propose to call this unemployment *involuntary*, though nothing of substance rides on the name.

The second ingredient is the hypothesis that workers care deeply about relative wages. This accomplishes two things. It rationalizes firms' decisions to pay efficiency-wage premia. And it explains why a worker laid off from a "good job" may prefer unemployment to a "bad job," at least for a while. The latter makes it possible for secondary labor markets to clear, or even to have excess demand, while involuntary (or socially useless) unemployment exists in primary labor markets.

Third, small costs of changing nominal wages and prices, coordination failures ("I'll cut my wage if you'll cut yours"), and notions of fairness<sup>15</sup> combine to prevent full adjustments to moderate shocks, whether nominal or real.

<sup>12</sup>Akerlof and Yellen actually assume what they call "near rationality." This is equivalent to rationality in the presence of fixed costs.

<sup>13</sup>A. Caplin and D. Spulber (1987) illustrate the idea; but their analysis pertains only to steady states with

constant  $s$  and  $S$ . In reality,  $s$  and  $S$  will undoubtedly be time varying. See, for example, the analysis in my 1981 article, or Avner Bar-Ilan and myself (1988).

<sup>14</sup>See Akerlof (1982) and Akerlof-Yellen (1987).

<sup>15</sup>See Daniel Kahneman, Jack Knetsch, and Richard Thaler (1986).

Consider what might happen in such a model if aggregate demand declines. Sales fall in many sectors of the economy, but unevenly. Although prices might drop in sectors experiencing extreme declines in demand, fixed costs keep most prices fixed. Virtually no wages fall due to firms' fully rational fears that wage cuts would lead to lower productivity, perhaps because wage cutting is widely perceived to be unfair.<sup>16</sup> Instead, most firms reduce output and employment.

The cyclically sensitive durable goods industries will be hit hardest by a typical downturn. It seems to be an interesting fact, which I will not attempt to explain, that they also pay very high wages.<sup>17</sup> Many of the workers laid off by those high-wage, cyclical industries will refuse low-status jobs in less cyclical industries, preferring to be unemployed steelworkers than employed hamburger flippers. Falling incomes lead to still-lower demand for goods, in a Keynesian multiplier process. Social welfare, I submit, falls.

### C. Third Challenge: Explain It Empirically

Economics is not an art form. So we must not be content with a coherent and vaguely sensible theory of unemployment—welcome as that would be. We must give the theory empirical content, test it, and estimate its central parameters.

In a sense, macroeconomics has progressed further on the empirical front than on the theoretical front. The truth of the matter is that empirical Keynesian models equipped with Phillips curves that allow for supply shocks have done rather well lately. Furthermore, the Phillips curve has been one of the strongest links in the empirical chain. Despite frequent reports of their demise, Robert J. Gordon's (1988) equations are alive

and well and living near Chicago. Academic economists jettisoned the Phillips curve not because of empirical failures, but because of a priori theoretical objections.<sup>18</sup> If we keep behaving like that, we may never become as useful as dentists.

What macroeconomics needs next is to give the new generation of Keynesian micro-foundations some empirical teeth. You can think of this as providing theoretical justification for the Phillips curve, if you wish. I prefer to think of it as providing empirical justification for all the theorizing.

### D. Fourth Challenge: *Devise Policies to Reduce It*

Logically, of course, this is the last step. But Keynes did not work that way, and the world will not wait while we perfect our models.

Observation of real economies suggests that the qualitative effects of demand management policies are more or less as taught in the elementary textbooks, or at least in *most* of them. Among other things, that means there will be an inflationary price to pay if unemployment is reduced by stimulating aggregate demand. It is the drive to subdue inflation, not any lack of knowledge about how to manipulate aggregate demand, that has accounted for high unemployment these past dozen years.

The nature of the policy challenge depends sensitively on whether or not the natural rate hypothesis is valid. If it is, then we can do no more than seek to flatten the short-run Phillips curve or reduce the natural rate by labor market policies. That remark is a place to begin a lecture, not to end one, so I will not pursue it further.

More enticing possibilities emerge if the natural rate is not so natural. Suppose, for example, that the equilibrium level of unemployment is strongly affected by hysteresis. Then a boost to demand might give the economy much more than a temporary high; it might actually lower unemployment *per-*

<sup>16</sup>Kahneman et al.; also Roger Kaufman (1984).

<sup>17</sup>In part, the high wages are compensation for the volatile employment. But I doubt that this is the whole story, for if low-wage, stable jobs were just as desirable as high-wage, variable ones, why would there always be queues of prospective workers at the high-wage firms?

<sup>18</sup>See my 1986 article; also Blanchard (1987).

*manently*. My Keynesian instincts tell me that the low-unemployment equilibrium must be better than the high-unemployment one.

The U.S. data for the 1980's look pretty consistent with the natural rate hypothesis to me—with a natural rate in the 5.5–6 percent range. But there is room for doubt. However, both the evidence of the senses and econometrics shun the natural rate hypothesis for Europe,<sup>19</sup> where none of the microeconomic factors comes close to explaining a quadrupling of unemployment. There a dose of expansionary policy might do the world a world of good.

### III. The Challenge to Microeconomists

Macroeconomics has long been regarded as the poor cousin of microeconomics, and with some justification. Surely it is mainly macroeconomists who have sullied the family name. But that is not because microeconomists have dealt with unemployment better; far from it. For the most part, microeconomic analysis ignores unemployment, as if it were an institutional detail of no great import.

Working within a full-employment framework would be justifiable on division-of-labor grounds if the premise of the neoclassical synthesis had been fulfilled. But plainly it has not been. Governments have failed to maintain anything like full employment and therefore have not created the conditions under which standard micro theory applies. Alternatively, the microeconomist's fixation on full-employment models might be legitimate if allocative decisions neither affected nor were affected by the overall level of employment. This might be true in some applications,<sup>20</sup> but there is no reason to think it holds generally. Let me illustrate with two examples.

<sup>19</sup>See Blanchard and Summers (1986).

<sup>20</sup>For example, some micro policies are too small to have meaningful macro effects (for example, airline deregulation). Another possibility is that central bank policy fixes real GNP.

### A. International Trade Theory

My first example is trade theory. Virtually all economists support free trade; but a frustratingly large number of noneconomists do not. Members of our fraternity are constantly amazed at the depth and strength of protectionist sentiment, which we view as evidence of either rent-seeking behavior or low intelligence. Doubtless, some protectionists qualify under both rubrics. But I want to suggest there is more to the matter.

One reason for economists' near-unanimous support of free trade is our use of the long-run, full-employment framework for policy evaluation. In our world, workers displaced by foreign competition move into industries in which our country has a comparative advantage. That can only raise productivity; so both GNP and social welfare should rise. How, except as viewed through the distorting lenses of a special pleader, could that be bad?

But people unencumbered by advanced degrees in economics see trade policy differently. They live in real space and time, where unemployment truly exists and workers displaced by foreign competition often move into unemployment rather than into new jobs. So they reason that our GNP will fall if our markets are opened to free trade. How, except in the strange world of the economic theorist, could that be good?

The two world-views generate rather different predictions. Which is right?

Consider a concrete example. Korean firms learn how to make television sets efficiently and want to export them to the United States. The TV industry and its workers petition Congress for a strict quota to "save jobs." Economists scoff at the idea. According to standard trade theory, America can only gain by opening its borders to Korean TVs. A quota cannot save jobs; it can only trap labor in an industry in which the United States has no comparative advantage.

Though oversimplified and missing many of the qualifications a good trade theorist would want, this conclusion probably characterizes the typical economist's view of the matter. And it is also probably the right view

for the long run. It might even be right for the short run, if the unemployment rate were 4 percent. But suppose Korea learns how to make TVs when the U.S. unemployment rate is 10 percent. Who can honestly assure a displaced factory worker that she will quickly find a new job at a wage close to her present one, as she would in the world envisioned by Ricardian comparative advantage? Isn't it more likely that she will suffer a spell of joblessness, perhaps a lengthy one? Aren't these short-run costs relevant to any social decision?

I anticipate your response and I agree with it: The appropriate solution is not to erect trade barriers but to pursue a vigorous full-employment policy so that displaced workers will be quickly reemployed. That is precisely my point. Conditions of full employment are necessary to validate standard propositions in trade theory. High unemployment calls many of these propositions into question. Both the positive predictions of trade theory and its normative prescriptions may be wrong. For example, Richard Brecher (1974) showed years ago that, when unemployment results from a rigid real wage, free trade may reduce both employment and welfare. Furthermore, if unemployment were eradicated by abolishing the wage floor, patterns of trade might reverse. Those who are wary of free trade may have a valid point in the presence of unemployment, as even Adam Smith realized (see H. Myint, 1958). At the very least, trade adjustment assistance should perhaps become a more integral part of the advocacy of free trade.<sup>21</sup>

Now, I am not trying to argue for protectionism. Though we may all be dead in the long run, *someone* will be alive. And a nation that protects one senile industry after another winds up looking like a nursing home for state capitalism. Economists correctly seek to avoid this outcome. Besides, the mere

existence of unemployment does not by itself imply that protection is better than free trade.

I *am* arguing, however, that trade theorists could do their job—the job Marshall wanted them to do—better if they paid more attention to the short run. At a minimum, it would narrow the communication gap between economists and the public. We insist on speaking in a long-run equilibrium dialect to people who live in a short-run disequilibrium world. No wonder what we say sounds Greek to them. We could, I believe, spend more time in their world without abandoning our own. And, if we did, everyone would benefit. Isn't that an unexploited Harberger triangle?

The phenomenon of unemployment, of course, is not unknown to trade theorists; and some interesting work has been done. But ask yourself what fraction of the enormous trade-theory literature deals with unemployment: 10 percent? 5 percent? Can that be an optimal allocation of resources?

### B. Public Finance

My second and last case in point, public finance, is a far greater offender. Once we get past the sizable literature on unemployment insurance, hardly any work in public finance even recognizes the existence of unemployment. Looking at this allocation of scholarly resources tempts me to prescribe a Pigouvian tax on full-employment theorizing.

Here is what we typically tell our youth about tax incidence. An excise tax is imposed on commodity *A*. In consequence, the price of *A* rises and the quantity falls. Resources released from the *A* industry migrate into the *B, C, D...* industries, where prices therefore fall and quantities rise. In the end, labor and capital are reallocated, relative prices adjust to the tax distortion, and another deadweight loss is born—about which we teach our students to worry deeply. Chances are that neither the price level nor aggregate employment ever arises.

Ordinary people may be forgiven for wondering if something important has not

<sup>21</sup>Michael Riordan and Robert Staiger (1987) show that trade adjustment assistance is welfare improving if terms of trade shocks are large enough. In their model, the unemployment results from informational asymmetries.

been left out of the story. Will displaced workers really be quickly reemployed in other industries? Aren't they more likely to experience a transitional period of joblessness, perhaps a long one? And won't the excise tax raise the price level? Old-fashioned macro-econometric models, you'll note, share this commonsense view. Plug an excise-tax hike into the DRI or Wharton model and you'll get back increases in both prices and unemployment. (You won't get the Harberger triangle, which is a shortcoming of these models.) Maybe, just maybe, the macro models are right and the micro theorists wrong.

I am not looking to score debater's points here. My claim is that many of the most cherished results of incidence theory change fundamentally once we allow for unemployment.

Consider, for example, the simple idea that an increase in an excise tax raises the price of the commodity to consumers. In one of the few papers on public finance theory in the presence of unemployment, Avinash Dixit (1976) showed that falling employment might so depress demand that the price of the taxed commodity actually falls.

Or consider what may be the most basic theorem of public finance: the irrelevance of the side of the market on which a tax is levied. We all have had fun explaining to our beginning students why it doesn't matter whether the payroll tax is levied on employers or employees. Then why, perhaps we should wonder, do Congress, labor, and management all think the decision so momentous? Sheer lack of understanding? Maybe. But maybe not.

I submit that part of the answer is, once again, that we economists insist on thinking long-run equilibrium while everyone else lives in short-run disequilibrium. The truth of the matter is that the incidence of the payroll tax may differ dramatically in the short and long runs; and, as Daniel Hamermesh (1980) showed with an empirically based simulation model, the short run may not be so short.

To see why, let us trace through what would happen if Congress abolished the employee's share and raised the employer's share by an equal amount—a nonevent in the eyes of conventional theory. Initially,

contractual wages are fixed, so both labor costs and take-home pay would rise. That, as we know, would create excess supply in labor markets, wages would fall, and so on. You can all complete the story leading to the conclusion that, in the end, nothing will have changed.

True enough. But the end is not the beginning. By blithely skipping over the adjustment period, we miss something important. Immediately after the law changes, firms are paying more and workers are receiving more; so capital bears the entire burden of the tax change—just as our mythical Congress intended. Had Congress shifted burdens in the opposite direction, labor would have lost out in the short run. So Congress's decision really does matter, at least for a while. No wonder workers and capitalists fret over where the tax is levied and are mystified by economists' indifference. We call them myopic. They call us out of touch. Both, I am afraid, are right.

Essentially this same point underlies an interesting recent paper by James Poterba, Julio Rotemberg, and Lawrence Summers (1986) which shows that a balanced-budget shift from direct to indirect taxation will reduce employment in a Keynesian model with nominal rigidities, but not in a classical full-employment model. In a similar vein, Th. van de Klundert and P. Peters (1986) coax a number of fascinating results out of a disequilibrium simulation model reminiscent of Dixit's theoretical paper. They find, for example, that a sales tax given back in a lump sum reduces employment dramatically more in the Keynesian first period than in the classical steady state.

Thus the differences between the long-run equilibrium results that we know and love (and teach to our young) and the short-run disequilibrium results that people actually experience are no mere quibbles. They may be fundamental. And that may be one reason why our advice so often falls on deaf ears.

Once again, the solution is not to abandon long-run analysis. The long-run questions are important and meaningful, and here economists are often right and the public wrong. Rather, the solution is to allow some consid-

eration of short-run employment effects to creep into and temper our analysis.

#### IV. Visions of Sugar Plums...

Lest I have failed to say anything provocative so far, let me conclude by trying once again, on this third night after Christmas, to get visions of New Jersey—or, if that is impossible, Massachusetts—dancing in your heads.

America now has a remarkable swath of prosperity in its northeast quadrant. It starts around Boston, runs through most of New England and down to the New York metropolitan area, then continues through central New Jersey and Philadelphia and on into Delaware, Maryland, and Washington, D.C., finally ending in portions of Virginia and North Carolina. By world standards, this is a very large economy; and, within it, unemployment rates of 4 percent and below are common.

Three questions cry out for answers. First, what created the boom? Second, how have these prosperous states managed to sustain such tight labor markets without blowing the lid off inflation? Third, could the entire United States accomplish something similar?

Both New Jersey and Massachusetts moved from the basket cases to showcases in a scant eight years.<sup>22</sup> New Jersey's unemployment rate went from 2.7 percentage points *above* the national rate in 1976 to 2.3 points *below* in 1984. Massachusetts' unemployment rate went from 2.6 points above the national rate to 2.7 points below between 1975 and 1983. How?

The answers are not well known and are probably not simple. Obviously, it was not Keynesian demand management by the state governments. However, rapid aggregate demand growth did play an important role in Massachusetts, which benefited from strong defense spending and "exports" of high-

technology manufactures to the rest of the United States. But New Jersey's economic renaissance came while its manufacturing sector was dwindling from 33 percent of private sector employment to only 23 percent. Services, especially information services, and construction led the way.

At the national level, we understand how to stimulate demand. So the more interesting question is how New Jersey and Massachusetts have kept inflation in check despite stunningly low unemployment rates.<sup>23</sup>

Two hypotheses can be ruled out immediately. The first is that stingy unemployment insurance and other tight-fisted government policies lowered the natural rates of unemployment. None of this is remotely close to the truth in either state.<sup>24</sup> The second hypothesis is that immigration or, alternatively, the use of guest workers provided these states with large influxes of labor at more or less fixed wages. No such thing happened. In fact, population growth in Massachusetts and New Jersey has been slower, and wage growth faster, than in the rest of the country. And I can assure you that New Jersey sends more guest workers—we call them commuters—to New York than New York sends to New Jersey.

The reasons must lie elsewhere. Let me offer two speculative possibilities. The first is hysteresis. Whether because outsiders became insiders, because high employment led to high capital formation, or because rapid growth stimulated innovation, the equilibrium unemployment rates in these two states may now be far lower than they were in 1975. If that is the explanation, we are left wondering whether the entire United States might do something similar.

The second has to do with openness. Each state of the union is a small open economy

<sup>23</sup> There is no CPI for New Jersey. But inflation rates in both the Philadelphia and New York City areas have been slightly *below* the national average. Inflation in the Boston area has run only slightly higher than national inflation.

<sup>24</sup> The taxpayers' revolt in Massachusetts is sometimes cited; but the timing is all wrong. Tax cuts began in 1981, but the "miracle" occurred between 1975 and 1983.

<sup>22</sup> On Massachusetts, see Ronald Ferguson and Helen Ladd (1986) and Katherine Bradbury and Lynn Browne (1987). The New Jersey story has been studied much less. See the New Jersey Economic Policy Council's *Annual Report* (1986).



with fixed exchange rates and no trade barriers vis-à-vis the others. It can therefore acquire the goods its citizens demand at more or less fixed prices in the huge national market. That is why textiles, shoes, refrigerators, and automobiles cost no more in New Jersey and Massachusetts than in the other 48 states. Nontraded goods, of course, are a different matter. Housing prices in the Boston and Princeton areas (indeed, in all the suburbs of New York), for example, are legendary. Were these states closed to trade with the rest of the country, their inflation rates would undoubtedly be much higher.

But what about the nation as a whole? America is certainly not a small economy. Nor is it as open to trade with the rest of the world as individual states are with the rest of the nation. Nor is the exchange rate fixed. So, if the national labor market tightened dramatically, we could not count on an infinitely elastic supply of imports to keep inflation as subdued as in New Jersey and Massachusetts. However, we *could* count on the world market to provide *some* moderation of inflationary pressures in tradable goods—at least as long as the rest of the world was not also in an exuberant boom. So perhaps the nation, with a balanced monetary and fiscal expansion and thorough-going free trade in goods (but not in labor), could emulate the Massachusetts and New Jersey success stories to some degree.

This is an important respect, I believe, in which free trade helps support a policy of low unemployment. And I argued earlier that low unemployment helps support free trade. That raises the tantalizing possibility of a virtuous circle in which high levels of aggregate demand create tight labor markets while open international trade moderates inflationary pressures. Now, that would truly be a grand neoclassical synthesis. But, to get there, policymakers, macroeconomists, and microeconomists all must rise to meet the challenge of high unemployment.

To do so effectively, we must leave the rubble of academic Star Wars behind us. We must stop arguing over easy questions with known answers (like whether socially useless unemployment exists), and start worry-

ing about difficult questions with unknown answers (like which of the theoretical explanations for unemployment are empirically important). Macroeconomics these last fifteen years has accomplished far too little that would make Alfred Marshall proud. It is time we gave that grand old man his due.

## REFERENCES

- Akerlof, George A., "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, November 1982, 92, 543-69.
- \_\_\_\_\_ and Yellen, Janet L., "A Near-Rational Model of the Business Cycle with Wage and Price Inertia," *Quarterly Journal of Economics*, 1985 Suppl., 823-38.
- \_\_\_\_\_ and \_\_\_\_\_, "The Fair Wage/Effort Hypothesis and Unemployment," unpublished, University of California-Berkeley, 1987.
- Altonji, Joseph and Ashenfelter, Orley, "Wage Movements and the Labour Market Equilibrium Hypothesis," *Economica*, August 1980, 47, 217-45.
- Ball, Laurence and Romer, David, (1987a) "Are Prices Too Sticky?," NBER Working Paper No. 2171, February 1987.
- \_\_\_\_\_ and \_\_\_\_\_ (1987b), "Real Rigidities and the Non-Neutrality of Money," June 1987, unpublished.
- Bar-Ilan, Avner and Blinder, Alan S., "Consumer Durables and the Optimality of Usually Doing Nothing," unpublished, January 1988.
- Bils, Mark J., "Real Wages over the Business Cycle: Evidence from Panel Data," *Journal of Political Economy*, August 1985, 93, 666-89.
- Blanchard, Olivier J., "Why Does Money Affect Output? A Survey," NBER Working Paper No. 2285, June 1987.
- \_\_\_\_\_ and Kiyotaki, Nobuhiro, "Monopolistic Competition and the Effects of Aggregate Demand," *American Economic Review*, September 1987, 77, 647-66.
- \_\_\_\_\_ and Summers, Lawrence H., "Hysteresis and the European Unemployment Problem," *NBER Macroeconomics Annual* 1986, 15-78.
- Blinder, Alan S., "Keynes After Lucas," *East-*

- ern Economic Journal*, July-September 1986, 12, 209-16.
- \_\_\_\_\_, "Retail Inventory Behavior and Business Fluctuations," *Brookings Papers on Economic Activity*, 2:1981, 443-505.
- Bradbury, Katharine L. and Browne, Lynn E., "The State of the New England Region," unpublished, Federal Reserve Bank of Boston, October 1987.
- Brecher, Richard A., "Minimum Wage Rates and the Pure Theory of International Trade," *Quarterly Journal of Economics*, February 1974, 88, 98-116.
- Caplin, Andrew and Spulber, Daniel, "Menu Costs and the Neutrality of Money," *Quarterly Journal of Economics*, November 1987, 102, 703-25.
- Clark, Kim B. and Summers, Lawrence H., "Labor Market Dynamics and Unemployment: A Reconsideration," *Brookings Papers on Economic Activity*, 1:1979, 13-72.
- Dixit, Avinash, "Public Finance in a Keynesian Temporary Equilibrium," *Journal of Economic Theory*, April 1976, 12, 242-58.
- Ferguson, Ronald F. and Ladd, Helen F., "Economic Performance and Economic Development Policy in Massachusetts," Discussion Paper D86-2, Harvard University, May 1986.
- Frank, Robert H., *Choosing the Right Pond*, New York: Oxford University Press, 1985.
- Geary, Patrick T. and Kennan, John, "The Employment-Real Wage Relationship: An International Study," *Journal of Political Economy*, August 1982, 90, 854-71.
- Gennard, John, "Job Security: Redundancy Arrangements and Practices in Selected OECD Countries," Paris: OECD, 1985.
- Gordon, Robert J., "The Role of Wages in the Inflation Process," *American Economic Review Proceedings*, May 1988, 78, 276-83.
- Grossman, Jean Baldwin, "The Response of Wages to the Minimum Wage: Theory and Empirical Evidence," unpublished doctoral dissertation, MIT, May 1980.
- Hamermesh, Daniel S., "Factor Market Dynamics and the Incidence of Taxes and Subsidies," *Quarterly Journal of Economics*, December 1980, 95, 751-764.
- Kahneman, Daniel, Knetsch, Jack L. and Thaler, R., "Fairness as a Constraint on Profit Seeking," *American Economic Review*, September 1986, 76, 728-41.
- Kaufman, Roger T., "On Wage Stickiness in Britain's Competitive Sector," *British Journal of Industrial Relations*, March 1984, 22, 101-12.
- Kiefer, Nicholas M. and Neumann, George R., "An Empirical Job-Search Model, with a Test of the Constant Reservation-Wage Hypothesis," *Journal of Political Economy*, February 1979, 87, 89-107.
- Mankiw, N. Gregory, "Small Menu Costs and Large Business Cycles: A Macroeconomic Model of Monopoly," *Quarterly Journal of Economics*, May 1985, 100, 529-37.
- Myint, Hla, "The 'Classical Theory' of International Trade and the Underdeveloped Countries," *Economic Journal*, June 1958, 68, 317-37.
- Pigou, A. C., *Memorials to Alfred Marshall*, London: Macmillan, 1925.
- Poterba, James M., Rotemberg, Julio J. and Summers, Lawrence H., "A Tax-Based Test for Nominal Rigidities," *American Economic Review*, September 1986, 76, 659-75.
- Riordan, Michael and Staiger, Robert W., "Sectoral Shocks and Structural Unemployment," unpublished, Stanford University, July 1987.
- Summers, Lawrence H., "Why Is the Unemployment Rate So Very High Near Full Employment?," *Brookings Papers on Economic Activity*, 2:1986, 339-96.
- Van de Klundert, Th. and Peters, P., "Tax Incidence in a Model with Perfect Foresight of Agents and Rationing in Markets," *Journal of Public Economics*, June 1986, 30, 37-59.
- New Jersey Economic Policy Council, *18th Annual Report*, Trenton, October 1986.

## THE POLITICAL ECONOMY OF TERRORISM<sup>†</sup>

### To Bargain or Not To Bargain: That Is The Question

By HARVEY E. LAPAN AND TODD SANDLER\*

In November 1986, news media revelations disclosed that the Reagan Administration had deviated significantly from its stated policy of never negotiating with terrorists when it traded arms to obtain the freedom of three Americans—Rev. Benjamin Weir in September 1985, Rev. Lawrence Jenco in July 1986, and David Jacobsen in November 1986. On January 26, 1987, terrorists, posing as policemen, kidnapped an Indian and three American professors at the American University of Beirut, thereby replacing the three Americans previously bartered away. Accepted wisdom, heard almost daily in newscasts, maintains that one should never bargain with terrorists since such negotiations encourage more hostage taking by making it a profitable activity; recent events in Beirut seem to support conventional views. Yet even the staunchest supporter of the no-negotiation strategy of precommitment, the Israelis, has made noteworthy exceptions in the case of the school children taken hostage at Maalot in May 1974,<sup>1</sup> and during the hijacking of TWA Flight 847 in June 1985. Another exception involved the Israelis' release of 1,150 Arab prisoners, including Kozo Okomato, in a negotiated swap for three Israeli soldiers in May 1985 (*The Economist*, 1987, p. 29). Okomato, a Japanese Red Army Faction member, was the sole surviving terrorist in the Lod Airport massacre of 1972, which left 27 people dead and 78 injured.

We use economic analysis in a simple game-theory framework to ascertain under what circumstances a government would want to precommit itself to a no-negotiation strategy. From the government viewpoint, we examine both the choice of deterrence expenditure (i.e., expense meant to reduce terrorist logistical success during incidents) and whether to negotiate or not.

Our analysis demonstrates that the beliefs and the resolve of the terrorists are crucial in identifying the rather restrictive scenarios in which a no-negotiation strategy is desirable in the case of a credible precommitment.<sup>2</sup> When governmental declarations are not completely credible and uncertainty characterizes the government's costs of not negotiating, then never negotiating is likely to be time inconsistent and not a plausible policy. In a multiperiod model, reputation effects may not be sufficient for a government to maintain a policy of never negotiating with hostage-taking terrorists owing to public choice considerations. Perhaps surprising, the conventional wisdom regarding the no-negotiation strategy does not withstand theoretical scrutiny except in a limited number of contrived cases.

#### I. Basic Structure of the Models

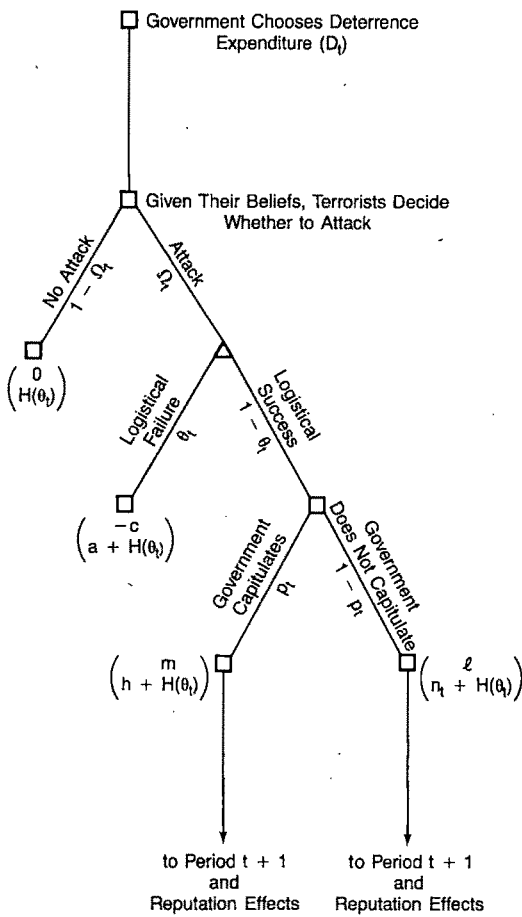
The analysis focuses on terrorist incidents that involve the taking of hostages (for example, skyjackings, kidnappings) for the purpose of gaining concessions (for example, ransoms, prisoners releases). There are two agents in the game—the terrorist group and the government; hostages are treated as exogenous participants. Initially, we present

<sup>†</sup>*Discussants:* Mancur Olson, University of Maryland; John Tschirhart, University of Wyoming; Benjamin Zycher, Rand Corporation.

\*Department of Economics, Iowa State University, Ames, IA 50011.

<sup>1</sup>Descriptions and details of transnational terrorist events between 1968 and 1979 are drawn from Edward Mickolus (1980).

<sup>2</sup>On terrorist rationality, see Scott Atkinson, Sandler, and John Tschirhart (1987).

FIGURE 1. GAME TREE—PERIOD  $t$ 

the underlying structure for a multiperiod sequential game that degenerates readily to a single-period game. In each period, the government has two potential strategic variables: how much to expend to deter an attack, and whether to negotiate or not in the event of a terrorist logistical success, whereby the terrorists manage to secure one or more hostages. The terrorists need to decide whether to attack. In Figure 1, the extensive form for this sequential game is displayed, complete with payoffs. First, the government chooses its  $t$ th period expenditures,  $D_t$ , to deter an attack; second, the terrorists decide whether to attack; and third, the government determines whether to negotiate in the event of a terrorist logistical success. The optimal strategy for each agent depends on its payoff

in each state and the beliefs that it holds as to the likelihood attached to each state. For any given state, the top number in the payoff vectors of Figure 1 denotes the terrorists' net benefit or cost, while the bottom number depicts the government's cost.

In period  $t$ , the terrorists receive 0 if they do not attack. In the event of an attack, three outcomes are possible: (i) the attack fails; (ii) the attack succeeds but the negotiation fails; and (iii) the attack and negotiation succeed. We assume that outcome  $i$  imposes a net cost of  $c$  on the terrorists, whereas outcomes  $ii$  and  $iii$  yield a net benefit of  $l$  and  $m$ , respectively, to the terrorists. The net benefit for a logistical success that does not produce concessions may be positive or negative. For a media-attracting skyjacking, publicity for a terrorist cause might make  $l$  positive even when concessions are zero. In the case of kidnappings, however,  $l$  is more apt to be negative, since the capture and subsequent maintenance of a hostage in a secret location is usually an expensive operation. The expected payoff ( $Z_t$ ) to the terrorists from an attack also depends upon the probability ( $\theta_t$ ) that the terrorists attack to a logistical failure and the probability ( $p_t$ ) that they attack to government capitulation.<sup>3</sup> Hence, from Figure 1, the terrorists' expected payoff from an attack is equal to

$$(1) \quad Z_t = (1 - \theta_t)[p_t m + (1 - p_t)l] - \theta_t c,$$

where  $m > l > -c$  and  $m > 0$ . The terrorists will attack whenever their expected payoff is positive; that is, whenever

$$(2) \quad c \leq c^* \equiv [(1 - \theta_t)/\theta_t] \times [p_t m + (1 - p_t)l].$$

From (2), the likelihood of an attack increases as either the probability of success ( $1 - \theta_t$ ) or the perceived likelihood of government capitulation ( $p_t$ ) increases. Equation (2) indicates that even a credible policy

<sup>3</sup>The time subscripts are introduced to allow subsequent generalization to a multiperiod setting.

of precommitment never to negotiation, which in turn implies that  $p_t = 0$ , may be insufficient to deter an attack if the terrorists derive net benefits from logistical success ( $l > 0$ ), from, say, publicity, even in the absence of concessions. The no-negotiation precommitment may also be insufficient to deter attacks when the cost associated with a logistical failure is low or negative.

In fact, terrorist groups that perceive benefits from logistical failure and logistical success ending in negotiation failure ( $l > -c > 0$ ) will attack regardless of a credible precommitment strategy. The Hezbollah, a pro-Iranian Shiite Fundamentalist terrorist group operating out of the Bekaa Valley in Lebanon, places a high value on martyrdom and could be placed in this category of groups. Not only does martyrdom give the victim a high perceived benefit, it assists the group to recruit. In a U.S. Department of State (1986, p. 19) report, the Hezbollah was said to hold many of the hostages taken in recent years in Beirut. Furthermore, the Hezbollah has been connected to the suicide bombings of the U.S. Marine barracks in October 1983, the U.S. Embassy in Beirut in April 1983, and the U.S. Embassy in Kuwait in December 1983. For the Hezbollah, the conventional wisdom regarding negotiations would not hold, since net benefits are derived under any outcome. A credible precommitment policy would, however, deter attacks if the group was solely motivated by concessions.

Turning to the government behavior, we denote its expenditure on deterring (and foiling) terrorist attacks as  $D_t$ . This expenditure will determine, along with nature, the terrorists' perceived probability of failure ( $\theta_t$ ) and will be incurred in all states. Government costs depicted in Figure 1 indicate that if no attack occurs, the government incurs no additional costs, but if an attack occurs and fails, the government incurs an additional cost of  $a$  ( $\geq 0$ ). If, however, an attack succeeds, the government must then decide whether to capitulate or not. The (current) cost of not capitulating is denoted by  $n_t$  and reflects the cost associated with hostage lives and resources expended. If the government does capitulate, it incurs a cur-

rent cost  $h$ , which may include perceived political cost and the cost associated with the consequences of freeing terrorists or augmenting terrorists' resources. Nevertheless, neither  $n_t$  nor  $h$  include *reputational* effects associated with negotiating at this juncture. In a true multiperiod setting, the government must consider how its negotiating behavior will affect its reputation by influencing the terrorists' perceived  $p_t$ , and hence the likelihood of an attack in the future. Moreover, a government must be concerned whether it can take current actions to alter terrorists' beliefs about the government's willingness to negotiate. One such posture that is often suggested is a *precommitment* policy never to negotiate, which we term credible provided that  $p_t = 0$  for all future  $t$ .

The government influences the terrorists' perceived failure rate through its expenditure on deterrence, in which

$$(3) \quad \theta_t = K(D_t) \quad \text{or} \quad D_t = K^{-1}(\theta_t) \equiv H(\theta_t),$$

where  $H(0) = 0$ ,  $\lim_{\theta_t \rightarrow 1} H(\theta_t) = \infty$ ,  $H' > 0$ , and  $H'' > 0$ . We further assume that the government does not know the resolve or fanaticism of the terrorists so that  $c \geq 0$  is a random variable with a probability density function of  $f(c)$ . From the government's perspective, the probability of an attack is the likelihood<sup>4</sup> that  $c < c^*$  (see (2)), that is

$$(4) \quad \text{prob}\{c < c^*\} = \Omega_t = \int_0^{c^*} f(c) dc.$$

Since  $c^*$  in (4) depends on terrorists' beliefs concerning capitulation and logistical success, the likelihood of an attack clearly depends on these beliefs, (i.e.,  $\Omega_t = F(\theta_t, p_t)$ ). By reducing the likelihood of a logistical success, increased deterrence expenditure would lower the perceived probability of an attack. Furthermore, an increase in terrorists' belief regarding government capitula-

<sup>4</sup>An alternative, but qualitatively similar, interpretation is to assume that there are many ( $N$ ) terrorist groups with the distribution of their values given by  $f(c)$ . Then  $\Omega_t$  represents the proportion of groups that attack;  $\Omega_t N$  is the number of such attacks.

tion encourages terrorist strikes. These results follow from partially differentiating (4) and the assumptions invoked thus far. Throughout, the two agents are assumed to have identical beliefs of the likelihood of attack.

## II. Single-Period Model: No Reputational Effects

In the single-period model, the government ignores reputation effects and takes as given the terrorists' belief,  $p_i$ , that the government will negotiate. If the negotiation decision is made *ex post*, the government would minimize its cost by negotiating if, and only if, capitulating is less costly (i.e.,  $n_i > h$ ). Hence, *ex post* cost in the event of a logistical success is  $\min(n_i, h)$ . From an *ex ante* perspective, expected cost to the government is

$$(5) \quad E[TC_i] = H(\theta_i) + \Omega_i \theta_i a + \Omega_i (1 - \theta_i) E[\min(n_i, h)],$$

where  $n_i$  is a random variable with a density function  $g(n)$ , and  $E[\cdot]$  is the expectations operator. The government determines its optimal level of deterrence by choosing  $\theta_i$ , *ex ante* to minimize (5). If  $h$  lies inside the range of  $n_i$ , then it is easy to show that  $E[\min(n_i, h)] < \min[E(n_i, h)]$ . This fact proves helpful when identifying costs associated with the precommitment strategy of never negotiating.

We denote  $\theta_i^*$  as the  $\text{argmin } E[TC_i]$  and  $TC_i^*$  as minimized expected cost. A simple comparative static analysis of the first-order conditions of (5) indicates that the optimal level of deterrence increases as (i) the likelihood of attack goes up, (ii) the expected cost of successful attacks rises, and (iii) the ability to deter attacks increases (i.e.,  $\Omega_i \theta_i = \partial \Omega_i / \partial \theta_i$  rises).<sup>5</sup> Furthermore, since the likelihood of attack is an increasing function of  $p_i$ , the optimal level of deterrence is also an increasing function of the government's be-

liefs of the terrorists' own beliefs concerning the government's willingness to capitulate. Finally, one might wonder where the terrorists' beliefs concerning capitulation are derived. If the terrorists know how the government behaves and if they further know the true distribution of the government cost from not capitulating, then consistency of expectations implies that  $p_i = \text{prob}[n_i > h]$ .

We are now prepared to examine the desirability of precommitment never to negotiate when deterrence expenditure is also a choice variable. In this case,  $\tilde{\Omega}$  denotes the *ex ante* probability of an attack when precommitment is credible ( $\tilde{\Omega} = F(\theta_i, 0)$ ). If the government adheres to its pledge, its expected costs are then

$$(6) \quad TC_i = H(\theta_i) + \tilde{\Omega}_i \theta_i a + \tilde{\Omega}_i (1 - \theta_i) E(n_i)$$

where  $E(n_i)$  indicates the expected cost of a successful attack, given that negotiations cannot occur. A comparison of (5) and (6) indicates that if a policy of precommitment eliminates all attacks (i.e.,  $\tilde{\Omega}_i = 0$  when evaluated at  $\theta_i^*$ ), then precommitment dominates the *ex post* decision. When, however, precommitment does not eliminate all attacks, precommitment would imply higher *ex post* costs from inflexibility in those incidents where costs would be minimized by capitulating (i.e.,  $E(n_i) > E[\min(n_i, h)]$ ). Thus, precommitment, even when credible, may not be optimal. In addition, when  $\tilde{\Omega}_i > 0$  and an attack occurs, the government may face a time consistency problem since holding firm to its policy may be more costly than capitulating. To compare the optimizing level of deterrence expense with and without precommitment, the first-order condition associated with (6) should be evaluated at the cost-minimizing deterrence level  $\theta_i^*$  for no precommitment. Such a comparison, while giving no definitive conclusions, implies that optimal deterrence expense under precommitment is apt to exceed that with no precommitment when either inflexibility costs are high or  $\tilde{\Omega}_i$  is near  $\Omega_i$  in value. In the latter case, precommitment does not significantly alter the likelihood of attack. Benefits from precommitment comes from its ability, if any, to change terrorists' beliefs,

<sup>5</sup>A more technical version of this paper, available upon request, contains the comparative statics details.

$p_t$ , and hence alter the probability of attack. Terrorist groups that do not believe the government's statement or its resolve will not be swayed. As long as  $n_t > h$  for some realizations, the time consistency problem can always surface as recent events in Lebanon have shown. Constitutional constraints or congressional hearings imposing huge perceived cost on those officeholders who capitulate may be the only means of raising  $h$  sufficiently to make precommitment time consistent.

### III. Multiperiod Models and Reputational Effects

In the context of hostage-taking incidents, reputational effects refer to the influence that the current government's negotiating behavior has on the beliefs of the terrorists concerning the government's willingness to grant concessions in the future. As before,  $p_t$  denotes the terrorists' beliefs concerning government capitulation in period  $t$ , and  $p_{t+1}$  represents their next-period beliefs. We assume the following updating behavior: (i)  $p_{t+1} = p_t$  if there is no opportunity to negotiate in period  $t$ ; (ii)  $p_{t+1} = p_{t+1}^N \leq p_t$  if the government refuses to capitulate after a successful attack; and (iii)  $p_{t+1} = p_{t+1}^E \geq p_t$  if the government capitulates in period  $t$ . Let  $J_{t+1}(p_{t+1})$  represent the (minimized) expected cost, from the current government's perspective, of an optimal program starting at  $t+1$ , with terrorist beliefs  $p_{t+1}$ . Since an increase in  $p$  augments the likelihood of attack, and therefore costs,  $dJ_{t+1}/dp_{t+1} > 0$ .

Public choice considerations are important when analyzing reputation effects. For example, governments that cannot or do not expect to be reelected would be unconcerned about reputation cost unless they are altruistic towards their successor. Even in the latter case, reputation may be nontransferable when terrorists do not believe that the current government's toughness will set the negotiation posture for the succeeding administration. To capture this aspect, we discount reputation costs by the probability  $\pi$  that the current government is in office in the ensuing period. The undiscounted future cost associated with a government capitula-

tion is

$$(7) \quad \Delta_{t+1}(p_t) = J_{t+1}(p_{t+1}^E) - J_{t+1}(p_{t+1}^N) \geq 0.$$

Given an attack, the government will negotiate *ex post* if, and only if,  $n_t > [h + \pi\delta\Delta_{t+1}]$ , where  $\delta$  denotes the natural discount rate. Deterrence expenditures are chosen *ex ante* to minimize<sup>6</sup>

$$(8) \quad TC_t = H(\theta_t) + \Omega_t\theta_t a + \delta[(1 - \Omega_t) + \Omega_t\theta_t][\pi J_{t+1}(p_t) + (1 - \pi)J_{t+1}(p_0)] + \Omega_t(1 - \theta_t) \times \{ \delta[\pi J_{t+1}(p_{t+1}^N) + (1 - \pi)J_{t+1}(p_0)] + E(\min[n_t, h + \pi\delta\Delta_{t+1}]) \},$$

where  $p_0$  is the reputation inherited by a new government. Equation (8) assumes the states of the world depicted in Figure 1, as well as the possibility of reelection or defeat for the government in the ensuing period. On the right-hand side of (8), the first term is the deterrence cost; the second is the additional expense (exclusive of reputation cost) to the government in the event of a terrorist attack failure; the third is the reputation cost in the event of no negotiation opportunities; and the fourth is the reputation costs in the situation of negotiation opportunities. If the government does not expect to be in office ( $\pi = 0$ ), the *ex post* negotiation rule and the cost-minimizing deterrence choice associated with (8) would be equivalent to the single-period case where reputation is unimportant.

Another case where reputation does not matter is that of exogenous expectations, or Nash-Cournot behavior, where  $p_{t+1} = p_t$  for

<sup>6</sup>Equation (8) also provides the recursive equation for determining  $J_{t+1}(\cdot)$ .

all  $t$ . That is, the terrorists do not use period  $t$  outcomes and observations to modify their beliefs about the government's willingness to capitulate. Since  $p_{t+1}$  is then independent of time  $t$  outcomes,  $\Delta_{t+1}$  is identically zero. Equation (8) then implies that the government's time-consistent solution is to negotiate if, and only if,  $n_t > h$ , as in the single-period model. The cost-minimizing choice of  $\theta_t$  is also unchanged from that of the single-period model.

This solution is not so naive since what really matters is not reputation per se, but rather the government's ability through its words and acts to alter that reputation. If terrorist groups believe that they have perfect information concerning  $h$  and the distribution of  $n_t$ , there is no reason for them to modify their beliefs due to time  $t$  events. If the terrorists are convinced that the government adheres to the time-consistent rule of capitulating when  $n_t > h$ , then, with perfect (perceived) information, the terrorists will set  $p_t = \text{prob}[n_t > h]$ . Moreover, under the circumstances, this will be the optimum negotiating rule for the government. Hence, we conclude that reputational effects are important *only under imperfect information*.

Nihilistic terrorist groups (for example, Japanese Red Army, Direct Action in France) have a strong distrust of the government's words and deeds and may well operate under the perception, false or otherwise, of perfect information. If such is the case, then, unless precommitment can eliminate attacks, and there is no assurance of that, the government faces time-inconsistency difficulties when they precommit.

When, however, terrorist groups learn so that their future beliefs are shaped by the government's current behavior, current capitulation by the government will increase future attacks ( $\Delta_{t+1} > 0$ ), thereby raising

costs from negotiations. With these reputation costs, the government will negotiate *ex post* only if  $n_t > [h + \pi\delta\Delta_{t+1}(p_t)]$ . Hence, the *ex ante* true probability that the government will negotiate is  $\sigma_t = \text{prob}[n_t > (h + \pi\delta\Delta_{t+1})]$ . Hence, an increase in  $\pi$  or  $\Delta_{t+1}$  will, *ceteris paribus*, decrease the likelihood of capitulation. Since reputation depends on terrorist beliefs,  $p_t$ , and the updating rule, the optimizing choice for deterrence depends upon initial conditions and the way in which terrorists learn or modify their beliefs. Thus, an analytical solution is not possible.

If the terrorists have access to the same information set as the government, both agents will know the true probability that the government will capitulate and  $p_t = \sigma_t$ . Since  $\sigma_t$  will then be independent of  $t-1$  events,  $\sigma_{t+1}$  (and hence  $p_{t+1}$ ) will, by induction, be independent of time  $t$  events. Under full information, there are consequently no reputation issues involved in the negotiation decision. The only consistent means of modeling reputation is to assume that the terrorists are uncertain about some aspect of the government information set. Strategic behavior then enters when subsequent events are used to modify the terrorists' priors.

## REFERENCES

- Atkinson, Scott E., Sandler, Todd and Tschirhart, John, "Terrorism in a Bargaining Framework," *Journal of Law and Economics*, April 1987, 30, 1-21.
- Mickolus, Edward F., *Transnational Terrorism: A Chronology of Events 1968-1979*, Westport: Greenwood Press, 1980.
- The Economist*, February 14, 1987, 302, 29.
- U.S. Department of State, *Patterns of Global Terrorism: 1985*, Washington: U.S. Department of State, 1986.



# Free Riding and Paid Riding in the Fight Against Terrorism

By DWIGHT R. LEE\*

Overt terrorist acts and the specter of terrorism impose significant costs on the community of civilized nations. These costs can be reduced by taking retaliatory action against terrorist organizations and the countries that sponsor them. Assuming a positive range over which retaliatory cost is less at the marginal than the resulting benefit from reduced terrorism, there exists some positive level of retaliation which is efficient from the perspective of the victimized countries.

An obstacle to achieving the efficient level of retaliation against terrorists results from the fact that much of the benefit from retaliation is general and cannot be captured entirely by the retaliating country. A fully efficient policy of retaliation will therefore require a cooperative response from all victimized countries. Achieving this cooperation, however, may confront the well-known free rider or prisoner's dilemma problem. While collectively the targets, and potential targets, of terrorism are better off retaliating, each will likely see its advantage best served by not retaliating. In this case it is quite possible that no one will retaliate when the best response is for everyone to retaliate.

A closer look at the problem of terrorism, however, suggests that the standard prisoner's dilemma may not be a significant obstacle to retaliation. Under plausible conditions, it will pay one country to retaliate even if other countries choose to free ride. Indeed, given the two options in the standard prisoner's dilemma setting, (cooperating (retaliating) or noncooperating (nonretaliating)), it is very likely that all victimized countries will retaliate; that no country will take a free ride. But this does not mean that the prospects for genuine cooperation in confronting terrorism are good. In the case

of retaliation against terrorist, free riding is only one way a country can increase the net benefit it receives from the public good provided by another. It is possible for one country to, in effect, "sell" the public good of reduced terrorism that is generated by the retaliation of another country. This paid-rider possibility has important, and obviously adverse, implications for the prospects that retaliation will result from a cooperative effort on the part of those countries victimized by terrorists.

In Section I, the implication of free riding and paid riding for retaliation against terrorists will be investigated in a two-country setting. In Section II, the circumstances that either encourage or discourage a country from being a paid rider on the retaliation of another country will be considered and some real world examples of paid riding will be discussed.

## I. Retaliation, Free Riding, and Paid Riding

The benefits that are generated by retaliation against terrorism are, at least in part, nonrival over countries. Terrorist activity imposes costs on people in all countries regardless of where a particular terrorist incident occurs. The fact that citizens of one country take precautions against the general threat of terrorism does not reduce the precautions that are advisable for citizens of other countries to take. Therefore, reducing the general threat of terrorism through retaliation generates a public good over many countries.

If the only benefit provided by retaliation was a public benefit, then the problem of motivating cooperative retaliation among countries would face the standard prisoner's dilemma. In the two-country case, each country will hope most to free ride on the retaliation of the other and fear most retaliating alone. The demands of individual rationality then result in neither country re-

\*Professor of Economics, University of Georgia, Athens, Georgia 30602.

taliating, which is collectively the worst possible outcome.

The likelihood of joint retaliation is increased by virtue of the fact that retaliation against terrorism provides country-specific benefits to the country doing the retaliating. A reduction in terrorism necessarily means fewer attacks against specific countries and their citizens. If a country can discourage terrorist attacks within its borders and against its citizens by retaliation, then differential (or country-specific) benefits accrue to a retaliating country. In this situation the importance of cooperation is reduced, since it is more likely that one country will retaliate regardless of what other countries do. It is a simple matter to construct a plausible  $2 \times 2$  payoff matrix that finds each country always choosing to retaliate.

There is another consideration that increases the likelihood that both countries will retaliate in the  $2 \times 2$  setting under discussion. When one country retaliates against terrorists, it generates both positive and negative externalities. As noted, such retaliation can be expected to reduce the overall level of terrorism and thereby provide general benefits to all countries. But, at the same time, the retaliation of one country will, by changing the relative cost to terrorists of alternative targets, tend to shift some terrorist activity to the other country.<sup>1</sup> Therefore the country-specific benefits one country receives from retaliation will increase with the retaliation of the other country. It is possible that even though country *B* would not be motivated to retaliate in the absence of country *A*'s retaliation, if country *A* does have the motivation to retaliate then so will country *B*.

Consider the  $2 \times 2$  payoff matrix in Figure 1. The left (right) column represents the decision to retaliate (don't retaliate) by country *A* and the top (bottom) row represents the decision to retaliate (don't retaliate) by *B*. The second number in each cell represents the payoff received by country *A* and the first number the payoff re-

		A	
		R	DR
B	R	120 150	100 125
	DR	95 130	105 95

FIGURE 1

ceived by country *B*. It is assumed that country *A* is being victimized by what it sees as domestic terrorists (terrorists whose grievances are primarily against country *A*) while country *B* is being victimized by what it sees as foreign terrorists (whose grievances are also primarily against country *A*). Note that it is to country *A*'s advantage to retaliate regardless of the decision made by country *B*. On the other hand, it is to country *B*'s advantage to retaliate only if *A* retaliates. Since *A* always follows a policy of retaliation the same policy will be followed by *B* and the equilibrium solution is for both countries to retaliate. The free-rider option is never attractive to *A* and therefore will be unattractive to *B*.

Consider, however, an additional option for country *B*. That option is in effect, to sell some of the terrorism reduction being provided by the retaliation of *A*. Exercising this paid-rider option requires offering the terrorist group safe haven in return for assurances that the terrorist activities against the accommodating country will be curtailed. By providing a sanctuary for terrorists, a country, in return for valued consideration, reduces the effectiveness of another country's retaliation at reducing the general level of terrorism. Free riding on the contribution of others does not reduce the general availability of the good being provided; paid riding does. Adding the paid-rider option into the situation described by the payoff matrix in Figure 1 can alter the solution dramatically.

In Figure 2 the Figure 1 payoff matrix has been expanded to include the "paid-rider" option for country *B*, the relevant payoffs being shown in the bottom row. While plausible that *A* will find the advantage in retaliating even though *B* free rides, it is

<sup>1</sup>See Todd Sandler, John Tschirhart, and Jon Cauley (1983) and Erik Im, Cauley, and Sandler (1987).

		A			
		R		DR	
B	R	120	150 <sup>1</sup>	100	125 <sup>2</sup>
	DR	95	130 <sup>3</sup>	105	95 <sup>4</sup>
	PR	140	75 <sup>5</sup>	115	80 <sup>6</sup>

FIGURE 2

equally plausible that *A* will cease to find the advantage in retaliating if *B* becomes a paid rider. At the same time, *B* can find the paid-rider option the most attractive regardless of *A*'s choice, even though when confined to the two options, retaliate or don't retaliate, the retaliation option is preferred when *A* retaliates. As shown in Figure 2, *B* will choose to be a paid rider, whether *A* retaliates or not. And with *B* acting as a paid rider, neutralizing much of the benefit generated by *A*'s retaliation, it no longer pays *A* to retaliate. By introducing the paid-rider option the equilibrium shifts from the best possible outcome (cell 1 in Figure 2) to the worst possible outcome (cell 6 in Figure 2).

## II. Further Implications and Illustrations

When considering a country's response to terrorism, it is plausible to argue that the greater the country-specific benefits a country receives from a reduction in terrorist activity against it, the greater the incentive that country has to retaliate. The paid-rider possibility, however, implies that this is not necessarily the case. The country that can realize great benefit from reducing terrorism through retaliation is a country that may also perceive great benefit from selling negative retaliation to terrorists. For reasons of location, wealth, influence, and history, certain European countries, for example, have been convenient places for terrorist attacks. The governments of these countries have not been blind to the possibility of reducing their cost from terrorism by offering foreign terrorists a sanctuary in return for restraint

in the host country. This possibility has been a compelling one in those countries (for example, France, Italy, Greece, and Cyprus) that have pursued such accommodations with foreign terrorist groups.<sup>2</sup>

There are no guarantees, of course, that terrorists will honor commitments made to paid riders. As the rash of terrorist bombings in Paris during 1986 and the hijacking of the Achille Lauro cruise ship in 1985 make clear, accommodations with foreign terrorists may not yield long-run advantage. However, a myopic government can be drawn into the paid-rider position by the short-run gains even though the long-run gains are negative. And once a myopic government has made an accommodation with terrorist groups, it can be difficult to extract itself from that accommodation when terrorists fail to honor their commitments. With the terrorists well entrenched because of past accommodations, the long-run advantage of reversing that accommodation can, from the perspective of a myopic government, be swamped by the short-run costs.<sup>3</sup>

The case of Israel is in some respects similar to that of the European countries just discussed. Israel would realize large benefits from a reduction in terrorism. In terms of my discussion, however, the distinguishing feature of Israel is that paid riding is not a realistic option. Even an extremely myopic Israeli government would see no advantage in seeking an accommodation with groups that refuse to recognize Israel as a nation and whose primary objective is its destruc-

<sup>2</sup> For a discussion of the tensions between the United States and Europe in general over Europe's more tolerant attitude toward terrorists (particularly Arab terrorists), see Christopher Hill (1986). France has harbored not only Arab terrorists, but for a long time provided sanctuary in southern France for Basque terrorists much to the displeasure of Spain. Cyprus and Italy have also accommodated Palestine terrorists in return for certain assurances.

<sup>3</sup> It should be noted, however, that France has recently taken some action against terrorists who have taken advantage of safe haven arrangements to carry out attacks in France. See Terrel Arnold and Neil Livingstone (1986, pp. 240-42).

tion.<sup>4</sup> With the paid-rider possibility precluded, the high benefits Israel receives from discouraging terrorism motivate a policy of significant retaliation.<sup>5</sup> The implication here is that a country may act as a paid rider with some terrorist groups while retaliating against others. Italy, for example, has retaliated with considerable success against the Italian-based Red Brigades. The Italian government is the primary target of the Red Brigades and therefore Italy is in much the same position with respect to the Red Brigades as Israel is with respect to the PLO. So while paid riding may appear to be an attractive option to a government when dealing with foreign terrorists whose grievances are not primarily with that government, paid riding will not be an attractive option to a government when dealing with domestic terrorists.

Retaliation against domestic terrorist groups provides clear private benefits from the perspective of the government whose country is being victimized. Maintaining its own commando unit to respond to particular terrorist incidents can provide the means for a country to focus its retaliation in such a way that it generates the most in terms of country-specific benefits. From the perspective of retaliating most efficiently against terrorism in general, and generating the greatest global benefits from retaliation, a strong case can be made for a transnational commando force. However, the temptation for a country to act as a paid rider with foreign terrorists while retaliating against domestic terrorists reduces the likelihood that a country will contribute to a transnational commando force. Not surprisingly, while a number of countries do have their own commando units, there exists no transnational commando forces.

In the standard public-good setting, free riders have no reason to object when others

contribute to the public good. To the contrary, free riders are obviously better off because of the contributions of others. Similarly, in the case of retaliation against terrorists, paid riders are made better off by the retaliation of others. The incentives created by paid riding, however, will motivate a paid rider to object to the retaliation of others, at least publicly. A country behaving as a paid rider will find it convenient to condemn another country's retaliation against a terrorist group that it (the paid rider) is accommodating while at the same time being pleased that the retaliation is taking place.<sup>6</sup> For example, much of the criticism the United States received for its attack against Libya came from governments that were reportedly pleased with the U.S. attack at the unofficial level.

### III. Conclusion

Terrorism is a complex phenomenon and no claim is made that a general theory of terrorism and retaliation against terrorists has been developed in the present paper. However, based on the plausible premises that 1) terrorists are rational actors and 2) retaliation against terrorists provides both country-specific benefits and general benefits over all countries, it is possible to argue that some retaliation against terrorists is desirable and ideally this retaliation will be a cooperative effort. The purpose of this paper is to consider obstacles to cooperative retaliation against terrorists.

On the one hand, the tendency for countries to contribute by retaliating against terrorists is stronger than would be true in a pure public-good setting because of the country-specific benefits that are generated by retaliation (plus the tendency for terrorists to substitute away from retaliating countries to nonretaliating countries). On the other hand, when considering retaliation against terrorism, an additional noncooper-

<sup>4</sup>In terms of my earlier discussion, Israel is being victimized by "domestic" terrorists. As such, Israel is represented by country *A* in Figure 2 and is not in a position to consider the paid-rider option.

<sup>5</sup>But, as suggested by the analysis in the previous section, if paid riding on Israel's retaliation became extensive enough, Israel's motivation to retaliate could erode.

<sup>6</sup>According to Hill, "In terms of domestic politics, ... there would still be pressures on them [European governments] to condemn the USA's use of force [against terrorists], even if it could be demonstrated that it was having a deterrent effect" (p. 96).

ative option exists that is not considered in the standard public-good discussion of free riding. That option is paid riding; that is, effectively selling the benefits from the retaliation of others by providing a sanctuary to terrorists. This paid-rider option can, from a country's perspective, dominate the retaliation option even though the retaliation option is preferred to the standard free-rider (don't retaliate) option.

The paid-rider option makes it likely that countries are retaliating against domestic terrorist groups while acting as paid riders with respect to foreign terrorists groups. An extremely noncooperative pattern of retaliating and paid riding can emerge in which the antiterrorism efforts of each country are being offset to one degree or another by the actions of other countries.

There are a number of examples of countries whose behavior toward certain terrorist groups fits quite closely the description of paid riding. These examples suggest that the paid-rider phenomenon deserves further

analysis as a means of better understanding the problem of terrorism.

## REFERENCES

- Arnold, Terrel E., and Livingstone, Neil C., "Fighting Back," in their *Fighting Back: Winning the War Against Terrorism*, Lexington: D. C. Heath, 1986, 229-47.
- Hill, Christopher, "The Political Dilemmas for Western Governments," in Lawrence Freedman et al., eds., *Terrorism and International Order*, Royal Institute of International Affairs, London: Routledge and Kegan Paul, 1986, 77-100.
- Im, Eric Iksoon, Cauley, Jon and Sandler, Todd, "Cycles and Substitutions in Terrorist Activities: A Spectral Approach," *Kyklos*, No. 2, 1987, 40, 223-55.
- Sandler, Todd, Tschirhart, John T. and Cauley, Jon, "A Theoretical Analysis of Transnational Terrorism," *American Political Science Review*, March 1983, 77, 36-54.

# Intervention Policy Analysis of Skyjackings and Other Terrorist Incidents

By JON CAULEY AND ERIC IKSOON IM\*

Terrorism represents a growing problem as illustrated by the statistics of the annual number of transnational terrorist events: In 1984, there were 597 such incidents; in 1985, there were 782; and in 1986, there were 848 (U.S. Department of State, 1986; *The Economist*, 1987, p. 36). International or transnational terrorism concerns activities involving terrorists or government participants from two or more nations. Incidents originating in one country and terminating in another are transnational, as are incidents involving the demands made of a nation other than the one where the incident is staged. The above statistics are especially bothersome, since one-third of the incidents resulted in casualties. Moreover, transnational terrorism does not include the vast number of domestic incidents whose effects do not cross political boundaries. In the face of this exigency, governments have tried to develop policies to thwart terrorism.

This paper applies the statistical technique of *intervention analysis*<sup>1</sup> to evaluate three specific antiterrorist policies: (i) increased airport security screening, (ii) increased security at U.S. embassies and other diplomatic missions, and (iii) the institution of the United Nations convention on preventing crimes against diplomatic personnel. The analysis here focuses on the substitution phenomenon, which poses a fundamental problem for counterterrorism efforts. This phenomenon indicates that terrorists will sub-

stitute out of one mode of attack (for example, aerial hijacking, barricade, and hostage taking) into another when government authorities crackdown on a particular mode. The substitution phenomenon, also known as transference in the criminal justice literature, was noted by Todd Sandler, John Tschirhart, and Cauley (1983) when modeling negotiations between terrorists and government authorities. Using a cross-spectral analysis, our paper with Sandler (1987) found empirical evidence for a substitution effect in terrorist activity. The analysis here goes a step further by testing the magnitude and the dynamic realization of the substitution effect. Policymakers are vitally concerned with the timing or pattern of the impact associated with a particular policy—that is, its dynamic realization.

## I. Theoretical Discussion

Terrorists are modeled as rational actors who adhere to the canonical, constrained optimization framework familiar to economists. Three general sets of terrorist choices can be delineated: the work-leisure choice, the legal-illegal decision, and the illegal-illegal choice. For our purposes, we assume that the first two choices have been made, and focus on the third, assuming that the individual has selected terrorism as a profession. In the simplest case, the final choice set would involve operating in one of two different modes of terrorist activity. Clearly, the specific choice between operative modes would depend upon their relative price structure.

When an antiterrorist policy is introduced, the initial impact is in the factor market where the marginal resource cost of the input is increased. The costs to the terrorists include the opportunity cost of time required to plan and execute the act. Moreover, the

\*Professor and Assistant Professor of Economics, Economics Department, University of Hawaii, Hilo, HI 96720-4091. We thank Will Gersch for his assistance in the initial stages of the project, and especially Todd Sandler for many helpful comments. Responsibility for any errors rests with us.

<sup>1</sup>Intervention analysis is also known as interrupted time-series analysis. For more details and references, see Richard McCleary and Richard Hay (1980).

likelihood of apprehension and conviction as well as the penalties expected must be included when estimating the costs associated with each terrorist mode of operation. If, *ceteris paribus*, government actions increase the likelihood of failure or the opportunity cost associated with a *given* mode of operation, terrorists would require larger expected remunerations in the factor market, and this in turn would result in a fall in the equilibrium level of employment of the input in the targeted mode. In the product market, this outcome generates a decrease in the supply of the impacted activity (for example, kidnappings) and an increase in the activity's price or cost to the terrorists. When more than one terrorist mode is examined, changes in relative costs will yield the standard substitution and income effects as terrorists adjust the composition of events in their campaign of terror accordingly. The substitution effect is especially problematic to policymakers since policy directed at thwarting one type of incident, say skyjackings, may induce terrorists to substitute into a now relatively less costly incident, say kidnapping. Although policy is effective in thwarting a specific type of terrorist event, the substitution effect may mean that it is not effective in curbing terrorism *per se*.

## II. Statistical Methods and Data

An appropriate approach to evaluating a policy impact on a time-series is "intervention analysis" (see McCleary-Hay). In general, intervention analysis provides a means of assessing the impact of a discrete intervention or interruption of a social process created, for example, by implementing a policy. Although William Landes (1978) employed more traditional techniques to ascertain the magnitude of an antiskysacking policy, only intervention analysis allows for an estimate of both the size and the time profile of the policy's impact.

Under the assumption that an observed time-series, such as that of a terrorist event, is one realization of a stochastic process, the first step of the procedure is to model the time-series,  $Y_t$ , and select the most appropriate ARIMA representation. Following

McCleary and Hay, an ARIMA model(s) is chosen for which the estimated coefficients for the preintervention periods are statistically significant, and the estimated residuals are not significantly different from white noise. If more than one model possesses the criteria, the Akaike Information Criterion is used to determine the best ARIMA structure. The next step is to model the impact of a specified policy by introducing an intervention component so that the time-series is depicted as consisting of two parts: the interruption or transfer component,  $f(I_t)$  and the noise component,  $N_t$ —that is,  $Y_t = f(I_t) + N_t$ , where  $I_t$  is a policy dummy which is zero for preintervention periods and unitary for postintervention periods. The policy impact in period  $t$  is

$$(1) \quad Y_t^* = Y_t - N_t = f(I_t),$$

where  $Y_t$  is the preintervention ARIMA representation for the time-series and  $f(I_t)$  is the so-called transfer function. The general expression of the transfer function used here is

$$(2) \quad f(I_t) = \frac{\omega_0}{1 - \delta_1 B - \delta_2 B^2} (1 - B)^D I_{t-p}$$

where  $B^n$  is a backshift operator which lags a variable by  $n$  periods. In (2),  $D$  is a binary variable which determines the impact status of the policy: if  $D = 0$ , the policy impact is permanent; if, however,  $D = 1$ , the impact is not permanent.<sup>2</sup> The lag indicator  $p$  denotes the number of periods by which the initial policy impact lags the policy intervention point. The first nonzero monthly impact is  $\omega_0$  when  $D = 0$  and approaches 0 when  $D = 1$ . The  $\delta_i$  parameters determine the stability of the model and the time pattern of impacts as shown in Table 1.

<sup>2</sup>This follows because  $(1 - B)I_{t-p} = 0$  prior to the impact's onset;  $= 1$  at the impact's onset; and  $= 0$  thereafter. When  $D = 1$ , the policy impact is of a transitory nature. The initial impact either dies out abruptly ( $\delta_1 = \delta_2 = 0$ ), or tapers off to zero ( $\delta_1$  and/or  $\delta_2 \neq 0$ ). This convergence is either gradual or oscillatory.

TABLE 1

	$\delta_1 = 0$	$0 < \delta_1 < 2$	$-2 < \delta_1 < 0$
$\delta_2 = 0$	abrupt	gradual	oscillatory
$0 < \delta_2 < 1$	oscillatory	oscillatory	oscillatory
$-1 < \delta_2 < 0$	oscillatory	oscillatory <sup>a</sup>	oscillatory

<sup>a</sup>Oscillatory only if  $\delta_1^2 + 4\delta_2 < 0$ ; otherwise gradual.

Stationarity of the model requires  $-1 < \delta_2 < 1$ ,  $\delta_1 + \delta_2 < 1$ , and  $\delta_2 - \delta_1 < 1$ . If the estimated values for  $\delta_1$  and/or  $\delta_2$  are statistically insignificant, then  $\delta_1$  and/or  $\delta_2$  are restricted to zero, and the impact model is reestimated until all the estimated coefficients are statistically significant following the rule of parsimony (McCleary-Hay). Although more complex transfer functions than (2) were tried, none passed the criteria used.

From (2), we have

$$(3) \quad Y_t^* (1 - \delta_1 B - \delta_2 B^2) = \omega_0 (1 - B)^D I_{t-p},$$

which implies

$$(4) \quad Y_t^* = \delta_1 Y_{t-1}^* + \delta_2 Y_{t-2}^* + \omega_0 I_{t-p}$$

when  $D = 0$ . Hence, (4) depicts the time pattern of impacts in terms of the parameters when the impact is permanent. By (4), the permanent impact asymptotically approaches  $\omega_0 / (1 - \delta_1 - \delta_2)$ .

Once an intervention model is estimated for a particular time-series of terrorist events using a specific intervention point, the model can then be reestimated using the same intervention point for a time-series consisting of all *other* incidents to detect a substitution effect. If the policy impacts for the two time-series are significant *and* opposite in direction, evidence of a substitution effect is present.

Data were drawn from Edward Mickolus' (1980) chronology of international terrorist events, which represents the only publicly available data set. Six time-series were extracted from Mickolus and included the monthly totals for the following: (i) skyjackings (*SJ*); (ii) all nonskyjackings (*NSJ*); (iii) barricade and hostage-taking events

(*BH*); (iv) all nonbarricade and hostage-taking events (*NBH*); (v) all terrorist acts directed against diplomats (*CD*); and (vi) all non-*CD* events (*NCD*).

For our empirical analysis, three policy intervention points were examined. On January 5, 1973, the U.S. government increased airport security by installing metal detectors in the hopes of curbing hijackings. Carry-on luggage and passengers were electronically screened for weapons. The second intervention point demarcated was January 1976 at which time embassy security was increased at U.S. missions. More specifically, "...Spending for embassy security increased from \$11.9 million in 1974 to 24.9 million in 1976."<sup>3</sup> Improvements in security at other likely targets for barricade and hostage-taking events were also made by the United States and other countries. The final intervention point was February 20, 1977, when President Carter proclaimed the U.S. adherence to the United Nations Convention on the Prevention and Punishment of Crimes against Internationally Protected Persons, Including Diplomatic Agents.<sup>4</sup>

### III. Empirical Results

As indicated above, testing the magnitude and form of the substitution effect between terrorist activities in response to an effective policy intervention involves interruptions of two different time series. The signs and estimated values for  $\omega_0$  and  $\delta_i$ ,  $p$ , and  $D$  indicate the policy effectiveness for the three intervention points. Table 2 indicates the statistically significant results at the .05 level; the three insignificant results are not reported.<sup>5</sup>

In the case of metal detectors, the January 1973 policy intervention was permanent ( $D = 0$ ). The estimated initial *monthly* im-

<sup>3</sup>From an April 21, 1987 letter to Todd Sandler from Jo L. Harben, Public Affairs Officer, Bureau of Diplomatic Security, U.S. Department of State.

<sup>4</sup>Yohan Alexander, Marjorie Ann Browne and Allen Nanes (1979, p. 78).

<sup>5</sup>Specific structure of the ARIMA models and other empirical results are available from the authors.



TABLE 2

	$\hat{\omega}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	$p$	$D$	$AI$
<i>SJ</i> <sup>a</sup>	-2.93	.58	-.67	0	0	-2.69
<i>NSJ</i> <sup>a</sup>	8.01	<i>R</i>	<i>R</i>	3	0	8.01
<i>BH</i> <sup>b</sup>	-1.03	.90	<i>R</i>	2	1	0

Note: *R* = Coefficient constrained to zero for its statistical insignificance in the preliminary estimation, and *AI* = Asymptotic Impact.

<sup>a</sup>January 1973.

<sup>b</sup>January 1976.

impact, measured by  $\hat{\omega}_0$ , is 2.93 fewer hijackings; the permanent impact was 2.69 fewer hijackings per month. The time path of the impacts is oscillatory since the estimated  $\delta$ 's satisfy  $\delta_1^2 + 4\delta_2 < 0$ . This finding is consistent with an attack-counterattack interaction between terrorists and officials as terrorists attempt to circumvent policies. The second row of Table 2 suggests the presence of a substitution effect, since the monthly total of nonskyjacking incidents increased by 8.01 after a three-month lag and remained thereafter ( $D = 0$ ). This lag may be due to the time required in switching from one mode to another.

These results suggest that the enhanced airport security has produced a mixed blessing: fewer skyjackings but an increase in other types of terrorist events. An overall evaluation of the policy's effectiveness would require a social comparison between skyjackings and other kinds of events. Without this comparison, one cannot say whether society is better or worse off with 2.93 fewer hijackings but 8.01 more nonhijacking events.

The results in row 3 of Table 2 indicate that the increase in embassy security had an abrupt but transitory influence ( $D = 1$ ) on the number of barricade and hostage taking events. The initial impact of 1.03 fewer events per month lagged the intervention point by 2 months. The initial impact gradually tapered off to zero, with most of the impact experienced in the first 21 months. A number of intervention models were tried for the non-barricade and hostage data, but  $\omega_0$  was con-

sistently insignificant; hence, there was no evidence of substitution.

The intervention analysis for the February 1977 UN convention did not show any significant impact on *CD*. The estimates of  $\omega_0$  were negative for a variety of transfer functions; however, none was statistically significant. Also, the concomitant impact on *NCD* was statistically insignificant. This result may be indicative of the "lack of teeth" in many UN policies as well as the difficulty in devising and implementing a multi-event thwarting counterterrorist policy.

To avoid a possible underspecification bias when estimating the coefficients of the transfer function for each of the time-series (i.e., *SJ*, *BH*, and *CD*), we simultaneously incorporated each of the three intervention points into the respective time-series. We found that each intervention point only influenced the targeted event's time-series; for example, the intervention point for metal detectors had a statistically insignificant effect on the time-series for *BH* and *CD*. This result, however, does not mean that significant substitution between and/or among these activities did not take place owing to other factors.

#### IV. Concluding Remarks

This paper highlights a critical problem confronting policymakers in their attempt to implement effective antiterrorist actions. Of the three antiterrorist policies investigated, only the installation of metal detectors was effective in both the short and long run. However, this policy was accompanied by a significant substitution effect that offsets, to some extent, the beneficial deterrence derived. These results suggest that an effective counterterrorist policy must increase the marginal resource cost of all terrorist modes of operation simultaneously. Only the UN convention was meant to thwart a number of terrorist modes of operation; but our empirical results showed that the policy was ineffective. In most instances, policy aimed at thwarting a wide range of terrorist events necessitates close cooperation and coordination among sovereign nations. The requisite cooperation has been difficult to achieve, given the high value nations place on their

autonomy. Substitution effects aside, the skyjacking crackdown was a unilateral policy of the United States in the beginning, and therein may be its narrowly defined success.

The barricade and hostage crackdown results provide evidence that, although an anti-terrorist policy may be effective in the short run, it may not be effective in the long run. Terrorists are clever and, if they do not substitute into alternative modes, they can devise means to circumvent a specific policy.

#### REFERENCES

- Alexander, Yonah, Browne, Marjorie Ann and Nanes, Allen S., *Control of Terrorism: International Documents*, New York: Crane, Russak and Company, 1979.
- Im, Eric Iksoon, Cauley, Jon and Sandler, Todd, "Cycles and Substitutions in Terrorist Activities: A Spectral Approach," *Kyklos*, No. 2, 1987, 40, 238-255.
- Landes, William M., "An Economic Study of U.S. Aircraft Hijacking, 1961-1976," *Journal of Law and Economics*, April 1978, 21, 1-31.
- McCleary, Richard and Hay, Richard A., *Applied Time Series Analysis for the Social Sciences*, Beverly Hills: Sage Publications, 1980.
- Mickolus, Edward F., *Transnational Terrorism: A Chronology of Events 1968-1979*, Westport: Greenwood, 1980.
- Sandler, Todd, Tschirhart, John T. and Cauley, Jon, "A Theoretical Analysis of Transnational Terrorism," *American Political Science Review*, March 1983, 77, 36-53.
- The Economist*, April 11, 1987, 303, 36.
- U.S. Department of State, *Patterns of Global Terrorism: 1985*, Washington: U.S. Department of State, 1986.

# THE NATURAL RATE THEORY RECONSIDERED<sup>†</sup>

## The Persistence of Unemployment

By ROBERT J. BARRO\*

In recent years, many economists have expressed concern about the persistence of unemployment (see, for example, Charles Bean, Richard Layard and Stephen Nickell, 1986, and Olivier Blanchard and Lawrence Summers, 1986). This study uses a time-series approach to assess the extent of this persistence in various countries over various time periods. Then I relate the measures of persistence to variables that matter theoretically for the economy's speed of adjustment. These variables include the extent of unionization, measures of the form of labor bargaining along the lines referred to by others as "corporatism," the size of the government, and the amount of governmentally mandated severance pay and notice for layoffs.

### I. Hall's 1979 Model of the Natural Unemployment Rate

The total labor force,  $L$ , consists of the unemployed,  $U$ , and the employed,  $E$ . I treat the labor force as constant, which means especially that I neglect cyclical variations in labor-force participation. Each unemployed person, treated as homogeneous, has a job-finding rate of  $f$  per year. Correspondingly, each employed person has a job-separation rate of  $s$  per year. With a constant labor force, the change over time in the unemployment rate,  $u_t = U_t/L$ , is given by

$$(1) \quad u_t - u_{t-1} = s(1 - u_{t-1}) - fu_{t-1}.$$

For constant values of  $s$  and  $f$ , this relation determines the steady-state (or "natural") unemployment rate as

$$(2) \quad \bar{u} = s/(s + f).$$

Hence the average unemployment rate increases with the ratio of  $s$  to  $f$ . Given a starting value  $u_0$ , the dynamics of  $u_t$  comes from solving equation (1) as a first-order difference equation to get

$$(3) \quad u_t = \bar{u} + (u_0 - \bar{u})(1 - s - f)^t, \quad t \geq 0.$$

Therefore  $u_t$  adjusts toward  $\bar{u}$  at a rate determined by the gross turnover rate,  $s + f$ . Equations (2) and (3) imply that equiproportionate increases in  $s$  and  $f$  leave  $\bar{u}$  unchanged, but raise the speed of adjustment from  $u_0$  to  $\bar{u}$ .

The model can be generalized to allow for fluctuations of  $s$  and  $f$  around stationary means,  $\bar{s}$  and  $\bar{f}$ . Then recessions correspond to periods where  $f$  is persistently below  $\bar{f}$  (or  $s$  is persistently above  $\bar{s}$ ). To approximate this kind of dynamics while retaining linearity, I instead add a shock,  $\varepsilon_t$ , to  $u_t$  in equation (1) to get

$$(4) \quad u_t = s + (1 - s - f)u_{t-1} + \varepsilon_t.$$

With  $s$  and  $f$  treated as constants, the persistence in the disturbance  $\varepsilon_t$  proxies for the persistence in separation and job-finding rates during recessions and booms.

In practice, heterogeneity for  $s$  and  $f$  in an economy may be important in accounting for job finding and job separation (see, for example, Kim Clark and Summers, 1979, and Michael Darby, John Haltiwanger and Mark Plant, 1985). But the introduction of these elements would not tend to disturb the main effect that I am focusing on. Namely,

<sup>†</sup>*Discussants:* Robert E. Lucas, Jr., University of Chicago; Edmund S. Phelps, Columbia University.

\*Harvard University, Cambridge, MA 02138. I am grateful for support of research by the National Science Foundation

economies with higher average values of  $s$  and  $f$  tend to have lower values for the autoregressive coefficient in equation (4).

## II. Measures of the Persistence of Unemployment

If  $\varepsilon_t$  is a moving-average process with  $k$  lags,  $MA(k)$ , then  $u_t$  in equation (4) is an  $ARMA(1, k)$  process. Table 1 shows estimates (from Micro TSP) of the  $AR1$  coefficient in this form,<sup>1</sup> using annual data on unemployment rates for 19 countries in the post-World War II period, 11 countries in the interwar period, and 3 countries prior to World War I. Since the data are annual averages, the  $MA$  terms would pick up time-averaging effects as well as the properties of  $\varepsilon_t$  in equation (4).

The concept of unemployment differs markedly across the countries and sample periods. In the post-World War II period, a few countries base unemployment on a labor-force survey, but most report registered numbers of unemployed, which ties in with the system of unemployment insurance. For earlier samples, the data are more limited and often refer to counts among union members.

To study persistence, it is unnecessary to compare levels of unemployment across countries or sample periods. In fact, a linear transformation of the data for each country/sample period would leave the results unchanged. Therefore, the persistence properties of a time-series would be robust to some, but not all, errors of measurement and concept.

My general procedure was to begin with an  $ARMA(1, 1)$  form, and then add enough additional  $MA$  terms to get the statistic for generalized serial correlation of residuals (for which I used the  $Q$ -statistic with 8 lags) below the 0.1 critical value. The first  $MA$

TABLE 1—VARIABLES FOR REGRESSIONS

Country/Sample	<i>AR1</i>	<i>UNION</i>	<i>CORP</i>
Australia, 1948–86	1.006 (.031)	.46	0
Austria, 1948–86	0.890 (.061)	.48	1
Belgium, 1948–86	0.877 (.049)	.72	1
Canada, 1948–86	0.881 (.073)	.25	0
Denmark, 1948–86	0.914 (.068)	.51	1
Finland, 1959–86	0.933 (.126)	.37	1
France, 1950–86	1.020 (.333)	.18	0
Germany, 1950–86	0.904 (.057)	.37	1
Ireland, 1948–86	0.980 (.059)	.31	0
Israel, 1950–86	0.653 (.108)	.66	1
Italy, 1948–86	0.927 (.055)	.32	0
Japan, 1949–86	0.937 (.069)	.19	0
Netherlands, 1949–86	0.913 (.057)	.32	1
New Zealand, 1949–86	1.034 (.054)	.37	0
Norway, 1948–86	0.954 (.115)	.48	1
Sweden, 1948–86	0.525 (.137)	.61	1
Switzerland, 1948–86	0.752 (.101)	.26	0
U.K., 1948–86	1.078 (.058)	.45	0
U.S., 1948–86	0.743 (.142)	.23	0
Australia, 1920–38	0.689 (.138)	.29	0
Belgium, 1922–38	0.802 (.104)	–	0
Canada, 1920–38	0.724 (.124)	.08	0
Denmark, 1920–38	0.682 (.209)	.24	0
Germany, 1920–38	0.715 (.223)	.13	0
Netherlands, 1920–38	0.778 (.098)	–	0
Norway, 1920–38	0.639 (.125)	.15	0
Sweden, 1920–38	0.390 (.315)	.19	0
Switzerland, 1925–38	0.819 (.147)	–	0
U.K., 1920–38	0.585 (.159)	.32	0
U.S., 1920–38	0.680 (.111)	.09	0
Germany, 1891–1913	0.462 (.431)	.10	0
U.K., 1891–1913	0.615 (.171)	.15	0
U.S., 1891–1913	0.609 (.177)	.05	0
U.K., 1852–1890	0.193 (.147)	–	0

Notes: *AR1* is from an  $ARMA(1, k)$  representation for the log of the unemployment rate. The standard error for the *AR1* coefficient is shown in parentheses. For New Zealand and Switzerland in the post-World War II period, the representation is for the level of the unemployment rate. *UNION* is the average over the sample period for the percentage of labor union membership in the labor force. *CORP* equals 1 for economies with predominantly an economywide structure of labor bargaining.

<sup>1</sup>I used the log-linear form, where  $\log(u_t)$  is the dependent variable, in most cases. This specification avoids the problem of having a functional form that allows  $u_t$  to become negative. However, the estimated measures of persistence are similar with a linear form, where  $u_t$  is the dependent variable.

term turned out to be sufficient in most cases.

For the post-World War II sample, 15 of the 19 point estimates for the *AR1* coefficient in Table 1 exceed 0.8. Most of the estimated values are insignificantly below 1.0 at the .05 level, according to the Dickey-Fuller test (W. A. Fuller, 1976, p. 373). However, many would be significantly below 1.0 at higher critical levels, and the joint hypothesis that all of the *AR1* coefficients equaled 1.0 would be rejected strongly in favor of the alternative that the coefficients were below 1.0. In any event, the general picture is one of high persistence of unemployment in the post-World War II period. The lowest estimated values for the *AR1* coefficients are .52 (*s.e.* = .14) for Sweden, .65 (.11) for Israel, .74 (.14) for the United States, and .75 (.10) for Switzerland.

For the earlier samples, the results show much less persistence of unemployment. For example, for the United Kingdom, the estimated *AR1* coefficient is 1.08 (.06) for 1948–86, .58 (.16) for 1920–38, .62 (.17) for 1891–1913, and .19 (.15) for 1852–90. For Germany, the values are .90 (.06) for 1950–86, .72 (.22) for 1920–38, and .46 (.43) for 1891–1913. However, for the United States, the pattern is more stable over time: .74 (.14) for 1948–86, .68 (.11) for 1920–38, and .61 (.18) for 1891–1913.

### III. Explaining Differences in the Persistence of Unemployment

I think of the estimated *AR1* coefficients in Table 1 as proxies for the term  $(1 - s - f)$  in equation (4). Therefore, country/sample periods with higher gross turnover,  $s + f$ , should have lower values for *AR1*.

Boyan Jovanovic (1979) provides a framework that I extend to analyze some determinants of  $s$  and  $f$ . Suppose that a randomly chosen job seeker, when attached to a randomly selected job, has the marginal product  $\theta$ , where the distribution,  $f(\theta)$ , is the same for all persons and jobs. *Ex ante*, workers and employers have imperfect information about the value of  $\theta$  in their particular match. After investing an optimally chosen amount of effort in examining each other, employment occurs if  $E(\theta) > \bar{\theta}$ , where  $\bar{\theta}$  is an opti-

mally chosen reservation value. The value of  $\bar{\theta}$  depends on the cost of search, the form of the distribution,  $f(\theta)$ , the value of time spent unemployed, and the possibilities for subsequent job separations. *Ex post*, the worker and employer acquire additional information about the value of  $\theta$ —in a simple case,  $\theta$  would be known precisely after some period on the job. Then, if  $\theta$  is below another reservation value,  $\hat{\theta}$ , a job separation occurs. The value of  $\hat{\theta}$  takes account of the costs of job turnover, the advantages of searching for a new job while unemployed rather than employed, etc.

In the context of European unemployment, it is natural to use the Jovanovic framework to assess the effects from restrictions on job loss. Suppose that restrictions from governments or unions make it costlier to discharge workers (by arbitrarily lowering the reservation value  $\bar{\theta}$ ). This type of restriction, if effective, would lower the job-separation rate,  $s$ . In the Jovanovic model, the knowledge that  $s$  has been reduced makes the initial reservation value  $\bar{\theta}$  higher than otherwise. The main point is that the initial job-match decision is more selective because of the constraint on subsequent discharge. Hence the restriction to a lower  $s$  leads to a lower job-finding rate,  $f$ . From the perspective of the Hall model, the reductions in  $s$  and  $f$  have an ambiguous effect on the average unemployment rate  $\bar{u}$  in equation (3). However, the persistence of unemployment rises, in the sense that  $(1 - s - f)$  increases in equation (4).

Economists often argue that unions restrict job turnover—in particular, they make the job-separation rate  $s$  inefficiently low, as sketched above. Presumably, such restrictions would not arise if the resulting inefficiencies could be internalized. The idea of corporatism, as discussed in Colin Crouch (1985), matters in this context. In a corporatist structure, labor-management bargaining occurs at something approaching an economywide level, with little power remaining for individual units. This setting may be undesirable in many respects, but it may be conducive to agreements that avoid some kinds of deadweight losses, such as those generated by inefficiently low labor turnover. Therefore, under these arrangements, the

persistence of unemployment would tend to be unrelated to variables such as the extent of union membership. On the other hand, competitive unions may generate deadweight losses from inefficiently low labor turnover. For example, restrictions on the discharge of workers may arise if unions mainly represent the existing employees (as in Assar Lindbeck and Dennis Snower, 1984), and if binding contracts cannot be entered into before employment starts. In this context the amount of restrictions on job turnover would tend to be increasing in the extent of union coverage. Hence I test the hypothesis that the persistence of unemployment increases with union coverage among economies that lack a corporatist bargaining structure.

Table 1 shows the variable *CORP*, which equals 1 for countries with predominantly a centralized structure of labor bargaining, and 0 otherwise (liberal). While this designation is subjective, I have followed the classifications presented by Crouch (pp. 115–18), with the modifications for Belgium and Switzerland proposed by Lars Calmfors and John Driffill (1987, pp. 18–21, and Table 3). Among countries not covered by Crouch, I designated Israel as corporatist, and all countries prior to World War II as liberal.

The variable *UNION* in Table 1 is an estimate for each sample period of the ratio of union membership to the labor force. In the post-World War II period, union coverage ranges from lows of 18 percent for France, 19 percent for Japan, and 23 percent for the United States, to highs of 72 percent for Belgium, 66 percent for Israel, and 61 percent for Sweden. For all countries on which earlier data are available, union coverage is substantially lower before World War II.

Blanchard and Summers (pp. 45–47), among others, note that effective union coverage can exceed union membership when the negotiated settlements apply also to nonmembers, as is the case in Germany and France. However, I lacked a way to quantify this idea for many countries, and therefore stuck with measures of union membership.

I looked first at the observations for which I had data on union coverage, along with the corporatist/liberal classification. This set comprises 30 of the 34 country observations

shown in Table 1: 19 in the post-World War II period, 8 in the interwar period, and 3 before World War I. The weighted<sup>2</sup> regression results for this sample are

$$(5) \quad AR1 = 1.05 \cdot CORP + 0.65 \cdot LIB$$

(.14)                      (.07)

$$- 0.36 \cdot CORP \cdot UNION + 0.86 \cdot LIB \cdot UNION,$$

(.27)                      (.25)

$$\text{weighted } R^2 = .38,$$

where *AR1*, *UNION*, and *CORP* are shown in Table 1,  $LIB \equiv 1 - CORP$ , and standard errors are in parentheses. The results suggest that unionization has a positive effect on the persistence of unemployment among the liberal economies (coefficient of .86, *s.e.* = .25), but no significant effect for the corporatist economies (–.36, *s.e.* = .27). Hence these results accord with the theoretical predictions. A test of equality for the intercept and slope coefficients across the corporatist/liberal division leads to rejection at less than the 1 percent level. Hence, the results suggest that the corporatist/liberal distinction is meaningful.

Since I lack data for the pre-World War II cases on some other variables, I now consider the sample of 19 post-World War II experiences. These cases provide less statistical support for effects from unionization or corporatism. The results that parallel those shown before are

$$(6) \quad AR1 = 1.00 \cdot CORP + .90 \cdot LIB$$

(.11)                      (.08)

$$- .25 \cdot CORP \cdot UNION + .27 \cdot LIB \cdot UNION,$$

(.22)                      (.27)

$$\text{weighted } R^2 = .38.$$

Hence the coefficients on the unionization variables are insignificant for the post-World War II sample.

<sup>2</sup>The procedure treats the *AR1* coefficients as observed with a measurement error, whose standard error equals the value shown in parentheses in Table 1. Then the true coefficient for case *i* satisfies a relation,  $AR1_i = a + b \cdot x_i + v_i$ , where  $x_i$  is a set of explanatory variables and  $v_i$  has the same distribution for all cases.

I then considered a measure of the size of government. The variable  $G$ , available for the post-World War II period, is the ratio for 1980 of government expenditures (total spending of consolidated general government) to GNP. I do not presently have data on this consolidated basis for countries before World War II. If  $G$  is a proxy for government regulations that inhibit labor turnover, then an increase in  $G$  leads to greater persistence of unemployment. However, if governmental regulations are substantially under the control of labor in a corporatist setting—where organized labor typically plays a strong role in the political process—then labor-market regulations that induce pure deadweight losses would be avoided. In this case the effect of  $G$  on the persistence of unemployment may be positive in a noncorporatist setting, but zero otherwise. Adding the variable  $G$  to the regression leads to

$$\begin{aligned}
 (7) \quad AR1 &= 1.31 \cdot CORP + .59 \cdot LIB \\
 &\quad (.28) \quad (.22) \\
 &- .22 \cdot CORP \cdot UNION + .29 \cdot LIB \cdot UNION \\
 &\quad (.20) \quad (.24) \\
 &- .65 \cdot CORP \cdot G + .80 \cdot LIB \cdot G, \\
 &\quad (.54) \quad (.49)
 \end{aligned}$$

weighted  $R^2 = .55$ .

The coefficient of  $G$  for the liberal group, .80,  $s.e. = .49$ , is significant at the .13 critical level.

I also examined Edward Lazear's (1987) measures of mandated severance pay and required notice for layoffs. These variables, available for a subset of countries in the post-World War II period, apply to most blue-collar workers with 10 years' experience. However, the severance and notice variables turned out not to have significant effects on the  $AR1$  coefficients.

With respect to unionization and corporatism, the stronger results for the larger sample that includes pre-World War II data can be understood from the following regres-

sion,

$$\begin{aligned}
 (8) \quad AR1 &= .99 \cdot CORP + .93 \cdot LIB \\
 &\quad (.09) \quad (.06) \\
 &- .23 \cdot CORP \cdot UNION + .18 \cdot LIB \cdot UNION \\
 &\quad (.17) \quad (.19) \\
 &- .30 \cdot DUM, \quad \text{weighted } R^2 = .68, \\
 &\quad (.06)
 \end{aligned}$$

where the variable  $DUM$  equals 1 for the pre-World War II cases and 0 otherwise. The results indicate that—holding fixed the dummy variable—unionization does not have a significant effect on the persistence of unemployment. On the other hand, the pre-World War II cases exhibit significantly lower persistence; the coefficient of  $DUM$  is  $-.30$ ,  $s.e. = .06$ . The general increase in unionization rates after World War II is one way to explain the generally higher persistence of unemployment in the recent period. It is this effect that accounts for much of the significance of unionization in equation (5). However, other variables, such as the size of government, that distinguish between the pre- and post-World War II periods would likely work as well.

#### IV. Conclusions

I presented a theoretical framework for assessing differences across economies in the persistence of unemployment. I also presented quantitative measures of persistence for a variety of countries in several sample periods. So far, there is only a little evidence about what causes the observed differences in persistence. Namely, there are suggestions that unionization and the size of government have positive effects on persistence among economies that lack a centralized structure of labor bargaining. Further research will attempt to clarify this relation and isolate other determinants of persistence.

#### REFERENCES

- Bean, Charles R., Layard, Richard and Nickell, Stephen J., "The Rise in Unemployment:

- A Multi-Country Study," *Economica*, supplement: 1986, S1-S22.
- Blanchard, Olivier J. and Summers, Lawrence H., "Hysteresis and the European Unemployment Problem," in Stanley Fischer, ed., *NBER Macroeconomics Annual 1986*, Cambridge: MIT Press, 1986.
- Calmfors, Lars and Driffill, John, "Centralization of Wage Bargaining and Macroeconomic Performance," unpublished, Institute for International Economic Studies, Stockholm, 1987.
- Clark, Kim B. and Summers, Lawrence H., "Labor Market Dynamics and Unemployment: A Reconsideration," *Brookings Papers on Economic Activity*, 1:1979, 13-60.
- Crouch, Colin, "Conditions for Trade Union Wage Restraint," in Leon N. Lindberg and Charles S. Maier, eds., *The Politics of Inflation and Economic Stagnation*, Washington: The Brookings Institution, 1985.
- Darby, Michael R., Haltiwanger, John C. and Plant, Mark W., "Unemployment Rate Dynamics and Persistent Unemployment under Rational Expectations," *American Economic Review*, September 1985, 75, 614-37.
- Fuller, W. A., *Introduction to Statistical Time Series*, New York: Wiley & Sons, 1976.
- Hall, Robert E., "A Theory of the Natural Unemployment Rate and the Duration of Unemployment," *Journal of Monetary Economics*, April 1979, 5, 153-70.
- Jovanovic, Boyan, "Job Matching and the Theory of Turnover," *Journal of Political Economy*, October 1979, 87, 972-90.
- Lazear, Edward P., "Job Security and Unemployment," unpublished, University of Chicago, May 1987.
- Lindbeck, Assar and Snower, Dennis, "Involuntary Unemployment as an Insider-Outsider Dilemma," in W. Backerman, ed., *Wage Rigidity and Unemployment*, London: Duckworth, 1984.



# Long-Term Unemployment and Macroeconomic Policy

By ASSAR LINDBECK AND DENNIS J. SNOWER\*

This paper bears a simple double message: when incumbent workers have some power in wage determination, then (i) there may be no natural rate of unemployment, and (ii) both supply-side and demand-side policies may have lasting effects on the unemployment rate. However, our analysis implies that demand-side policies in the product market may be much less reliable, and operate through more complex channels, than the traditional Keynesians envisaged.

To study the consequences of demand-management policies for the labor market, we need to explore the transmission of product-demand shocks to the labor market. Without denying the practical importance of sluggish wages and prices in this transmission process over the short run, we here set out to examine the effectiveness of macroeconomic policies when wages and prices are flexible, in the sense that agents set them freely in response to policy changes. In this context, as we shall see, there are transmission mechanisms which permit both pro- and countercyclical movements of real wages.

We assume that pricing, production, and employment decisions are made by imperfectly competitive firms (taking wages as given), and that nominal wages are set by workers (who take the effect of wages on employment into account). (The substance of our argument would remain unchanged if nominal wages were determined through negotiations between firms and workers.) The firms' decisions yield a relation between the real wage and aggregate labor demand—the "labor-demand relation," for short. The wage setters' only target variables are assumed to be the real wage and employment, and thus the wage setting in effect determines a point

on the labor demand relation (i.e., a real wage and a level of employment).

As we have no quarrel with transmission mechanisms by way of changes in the real wage and concomitant movements along the labor-demand relation, we concentrate here on the ways in which macroeconomic policies may affect wages and employment through shifts in the labor demand relation. We proceed in two steps. In Section I, we inquire how such policies change the relation between real wages and labor demand. In Section II, given a change in this relation, we examine how wages, employment, and unemployment are determined.

## I. Transmission of Macroeconomic Policies to the Labor Market

We represent a firm's demand function by

$$(1) \quad P = P(Q, A), \quad P_1 < 0, P_2 > 0,$$

where  $P$  is the price,  $Q$  is product demand, and  $A$  is a shift parameter, which may be varied through demand management policies. Moreover, let the firm's production function be

$$(2) \quad Q = f(L), \quad f_1 > 0, f_{11} < 0,$$

where  $L$  is labor.

Suppose that each firm, when maximizing its profit subject to its product-demand function and production function, takes the nominal wage ( $W$ ) as given, so that the real marginal value product of labor is equal to the real wage:

$$(3) \quad b \cdot f_1 = W/P,$$

where  $b = (1 - (1/\epsilon))$  and  $\epsilon$  is the price elasticity of the firm's product-demand function.

Assuming (merely for simplicity) that there is a given number ( $M$ ) of identical firms in the economy and that their product-demand

\*Institute for International Economic Studies, University of Stockholm, S106 91 Stockholm, Sweden, and Birkbeck College, University of London, 7 Gresse Street, London W1PA 1PA, England, respectively.

functions are independent of one another, the aggregate labor-demand relation is

$$(4) \quad N = M \cdot L = M \cdot L(W/(b \cdot P)),$$

$$L = (f_1)^{-1} \quad \text{and} \quad L' < 0.$$

This simple condition tells us that, under the imperfectly competitive conditions outlined above, demand-management policies can shift the aggregate labor-demand relation (equation (4)) only if such policies are able to change one or more of the following three variables: (a) the number of firms in the economy ( $M$ ), (b) the marginal product of labor ( $f_1$ ), or (c) the price elasticity of product demand ( $\varepsilon = 1/(1 - b)$ ).

It should be noted that the labor-demand relation does *not* depend directly on the shift parameter ( $A$ ) of the product-demand functions. Thus, a policy which merely shifts the product-demand functions (without affecting any of the variables above) leaves the aggregate labor-demand relation unchanged.

Of the three variables above, the demand elasticity is probably not a reliable and systematic channel for the transmission of policy shocks from the product to the labor market. There do not appear to be compelling reasons to believe that this elasticity rises (falls) systematically whenever product demand rises (falls).

As for the other two channels of transmission, expansionary demand management policy may (a) create incentives for the entry of new firms (which in turn raises the demand for labor associated with any given real wage), and/or (b) raise the marginal product of labor—either *directly*, by government policies which augment the industrial infrastructure of the economy, or *indirectly*, when the policy leads to a rise in the use of factors which are complementary to labor or to a fall in the use of substitutes for labor.

The latter, indirect effect on the marginal product of labor may have a significant role to play when there is excess capital capacity and the product-demand stimulus raises firms' rate of capital utilization. In that event, workers are simply recalled to operate unmanned machines and reestablish existing assembly lines. The point is that the plant

and equipment which is brought into use in the course of cyclical upswings is usually complementary to labor, and this means that the rise in the capital utilization rate may be expected to raise the marginal product of labor.

In short, under flexible wages and prices set by imperfectly competitive agents, our analysis leads us to identify one short-run, one medium-run, and one long-run channel whereby these shocks may shift the aggregate labor-demand relation. The short-run channel involves a change in the rate of capital utilization; the medium-run channel operates through the entry and exit of firms; and the long-run channel works via the buildup and rundown of industrial infrastructure. (For a detailed analysis of these channels, see our 1987c paper.)

What are the policy implications of these lines of thought? First, the short-run transmission mechanism, involving changes in the rate of capital capacity utilization, is operative only as long as there is excess capital capacity—regardless of the rate of unemployment. Thus, demand-management policies may be able to raise employment at constant (or even rising) real wages when there is excess capacity, but unable to do so at full capacity utilization. Second, the removal of barriers to the entry of firms may be an important ingredient in making demand-management policy effective. Third, changes in government expenditure on industrial infrastructure may have a much larger impact on the labor market, at least in a long-run perspective, than have spending changes on goods which are not complementary to labor (as in the case of tax reductions, increased transfer payments, or greater government purchases of consumer goods).

## II. The Labor Market

Having examined the effect of demand-management policies on the relation between the real wage and aggregate labor demand, we now turn to the determination of a wage-employment point on this relation and to the associated level of unemployment.

In particular, we show that if incumbent workers have some market power in the

negotiations over nominal wages, then policy-induced shifts in the aggregate labor-demand relation may give rise to persistent changes in the level of unemployment. In this context, there is no natural rate of unemployment as commonly envisaged by natural rate theories. In other words, when wage-price expectations are correct, unemployment is not necessarily at a unique rate, determined exclusively by the tastes, technologies, and endowments of the agents in the economy.

Since we wish to focus our attention on how the exercise of market power by incumbent workers may be responsible for persistent effects of macroeconomic policies on unemployment, we begin by considering the source of incumbent market power. In line with the insider-outsider theory (see, for example, our 1987a paper), we identify labor turnover costs as the source. These costs may take a wide variety of forms, for example, costs of hiring and firing, costs arising out of differences in cooperation and harassment activities among incumbents and new entrants, and costs due to the effect of labor turnover on work effort. These costs give the incumbent workers ("insiders") the ability to hurt their employers when there is disagreement in wage negotiations, that is, the turnover costs provide threat points in the wage negotiation process. When the insiders have market power, their employers cannot entirely pass the turnover costs onto them in the form of correspondingly lower wages.

Consequently, the insiders are able to negotiate their wages without fully taking account of the interests of the unemployed workers ("outsiders") and the newly hired workers ("entrants"). However, after an outsider is hired, he is assumed to remain an entrant only for a limited span of time, which is sufficient for the entrant wage contract to expire and for the worker to become associated with the insiders' labor turnover costs. At the end of this time span, the entrant turns into an insider.

Modifying the firm's marginal productivity condition (3) to include the employment of insiders ( $L_I$ ) and entrants ( $L_E$ ), we get

$$(5) \quad b \cdot f_i(L_I, L_E) = W_i/P; \quad i = I, E,$$

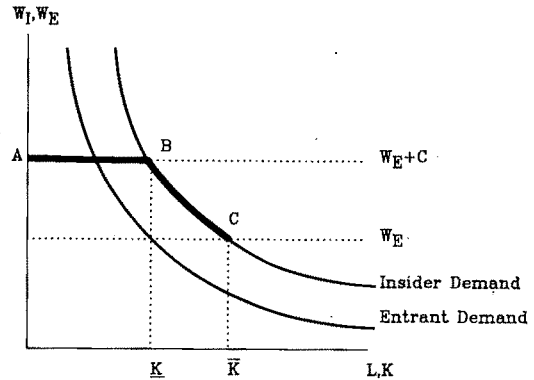


FIGURE 1. THE FIRM'S EQUILIBRIA

where  $W_I$  and  $W_E$  are the nominal wages of insiders and entrants, respectively, and  $f_I$  and  $f_E$  are their marginal products adjusted for the relevant labor turnover costs. For instance,  $f_I$  could be the insiders' marginal product plus their marginal firing cost and  $f_E$  could be the entrants' marginal product minus their marginal hiring cost.

The labor-demand relations for insiders and entrants are illustrated by the downward-sloping curves in Figure 1. In particular, let  $K$  be the firm's incumbent workforce and suppose that the insider wage is set so that the firm never has the incentive to replace incumbents by entrants. Thus the insider demand relation is  $P \cdot b \cdot f_I(L_I, 0) = W_I$  and the entrant demand relation is  $P \cdot b \cdot f_E(K, L_E) = W_E$ .

Turning to wage determination, our analysis requires that the insider wage be the outcome of negotiations between each firm and its insiders (who may bargain collectively or individually), and that the insiders have some market power in these negotiations. Yet, merely for expositional simplicity, we assume that the insiders have complete market power in the determination of the nominal insider wage and that each insider sets his wage "individualistically" (taking the wages and employment of the other insiders as exogenously given), so that each insider views himself as the marginal employee in his firm.

Then the nominal insider wage  $W_I$  will be set as high as possible, subject to the con-



These three conditions are important for the following reasons. First, it is obvious that the insider wage will respond to shocks only if the shocks are observed prior to the wage decision. If, on the contrary, the shocks are observed afterwards, then our model generates employment fluctuations at constant insider wages. Second, insiders' influence over turnover costs may give them the ability to prevent underbidding by laid-off workers. They may do so by refusing to cooperate with the underbidders (thereby reducing their productivity), by creating a hostile work environment for the underbidders (thereby raising their reservation wage), or by threatening to strike or work-to-rule. (See our 1988 paper.) Third, the existence of a seniority system permits the insiders to identify in advance the laid-off workers whose underbidding activities are to be thwarted.

Let us consider the effects of supply-side shocks. Suppose that these shocks are anticipated in the wage decisions, that insiders can influence labor turnover costs, and that a seniority system exists. Let the initial labor market equilibrium be given by Point  $e_1$  in Figure 2 (where the incumbent workforce lies in the range  $\underline{K} \leq K \leq \bar{K}$ ). Thereupon an unfavorable supply-side shock occurs, which shifts the labor market equilibrium locus from  $DEF$  to  $DE'F'$ . The insider wage may fail to fall in response to this shock, even though workers are laid off. The reason is that if the laid-off workers should try to regain their jobs by offering to work for a lower wage, the remaining insiders could prevent this from happening by manipulating the labor turnover costs (for example, by harrasing the underbidders.) Consequently, the labor-market equilibrium moves from Point  $e_1$  to  $e_2$ .

Now suppose that, later on, a favorable supply-side shock occurs, shifting the labor market equilibrium locus back out to  $DEF$ . Now the insiders have the opportunity to raise their wage without fear of being displaced by other workers. As result, the insider wage rises and employment remains unchanged. The labor-market equilibrium moves from point  $e_2$  to point  $E$ .

As we can see, when the incumbent workforce lies in the range  $\underline{K} \leq K \leq \bar{K}$ , favorable and unfavorable supply-side shocks do not

have symmetric effects on wages and employment: the unfavorable shock reduces employment, but the favorable shock does not increase employment. (If we instead assume that *both* insiders *and* firms have power over the insider wage, then the unfavorable shock reduces employment merely by more than the favorable shock increases it). We call this phenomenon "asymmetric persistence" of supply-side policy effects.

Thus, a succession of downward and upward shifts of the equilibrium locus yields a wage-employment ratchet, characterized by an upward trend in wages and a downward trend in employment. This ratchet disappears once the insider wage reaches the level  $W_E + C$ . The reason is that the insiders cannot raise their wage above this level, for otherwise they would be replaced by outsiders. At  $W_I = W_E + C$ , upward and downward shifts of the equilibrium locus lead to variations in employment at constant real wages. (This is illustrated by the arrows between equilibrium points  $E'$  and  $E$  in Figure 2). Here, there is "symmetric persistence" of supply-side policy effects. (Other models of symmetric persistence are contained in Olivier Blanchard and Laurence Summers, 1986; N. Gottfries and H. Horn, 1987; and ourselves, 1987b.)

Note that when there is no explicit or implicit seniority system or when insiders cannot influence turnover costs, then the insiders will be unwilling or unable to prevent underbidding from occurring. Consequently, favorable and unfavorable supply-side shocks lead to variations of the insider wage at constant employment.

Now turn to the effects of demand-side macroeconomic policies on the labor market, in the light of the discussion of the transmission mechanisms in Section I. We consider the three demand-side transmission mechanisms of Section I in turn. First, some types of government investment in industrial infrastructure will raise the marginal product of labor and thereby shift the labor-market equilibrium locus outwards. Conversely, a rundown of infrastructure causes the locus to shift inwards. The resulting effects on wages, employment and unemployment are basically the same as the effects of the supply-side policies considered above.

Second, demand-side policies which lead to the entry of new firms serve to raise employment of entrant workers, who receive the reservation wage (provided that union agreements or government legislation do not prevent new firms from hiring labor at the reservation wage). After these workers turn into insiders, they receive the insider wage. (For instance, letting the firm in Figure 1 be a new firm,  $K$  entrants are hired at wage  $W_E$ , and once they achieve insider status, their wage becomes  $W_I = W_E + C$ .)

Finally, consider demand-side policies which raise the marginal product of capital by increasing the rate of capital utilization. Assuming that the capital brought back into operation is complementary with labor, the insider and entrant labor-demand curves (in contrast to those pictured in Figure 1) may be upward sloping at cyclically low levels of capital capacity utilization and downward sloping only at full capacity utilization. Accordingly, the labor-market equilibrium locus (in contrast to that pictured in Figure 2) may have both upward- and downward-sloping portions. This means that the demand-side policies above can move the labor-market equilibrium point along either an upward- or a downward-sloping labor market equilibrium locus. (See our 1987c paper.)

#### IV. Concluding Remarks

Our analysis suggests that the entry and exit of firms may play an important long-term role in the transmission of product market shocks to the labor market. In this light, lower barriers to entry by firms in the United States than in Western Europe may help explain why U.S. employment recovered more rapidly from the recession of the late 1970's and early 1980's than European employment did.

We also argue that demand management policies which have "supply-side" effects on labor productivity—for example, policies which stimulate the rate of capital utilization or expenditures on industrial infrastructure (such as that undertaken by Western governments in the 1950's)—may have a larger impact on employment than policies without

such supply-side effects (such as the transfer payments which have commanded progressively larger portions of European government budgets in the postwar period).

Finally, our analysis suggests that aggregate supply shocks may affect the labor market more directly and speedily than most aggregate demand shocks do. In this light, it appears that the overall level of unemployment in Europe during the 1950's and 1960's may have been low partly on account of the steady stream of expansionary supply-side shocks (such as a falling real price of oil). By contrast, European unemployment may have been comparatively high since the mid-1970's because the contractionary supply-side influences (including the overshooting of product wages) may have been difficult to counteract through demand-management policies, particularly in the face of limited entry of firms and insufficient excess capital capacity.

#### REFERENCES

- Blanchard, O. and Summers, L., "Hysteresis and the European Unemployment Problem", in S. Fischer, ed., *NBER Macroeconomics Annual 1986*, Cambridge: MIT Press, 1986, 15–78.
- Gottfries, N. and Horn, H., "Wage Formation and the Persistence of Unemployment," *Economic Journal*, December 1987, 97, 877–84.
- Lindbeck, A. and Snower, D. J., (1987a) "Efficiency Wages versus Insiders and Outsiders," *European Economic Review*, February 1987, 31, 407–16.
- \_\_\_\_ and \_\_\_\_\_, (1987b) "Union Activity, Unemployment Persistence and Wage-Employment Ratchets," *European Economic Review*, February 1987, 31, 157–67.
- \_\_\_\_ and \_\_\_\_\_, (1987c) "Transmission Mechanisms from the Product to the Labour Market," Seminar Paper, Institute for International Economic Studies, University of Stockholm, 1987.
- \_\_\_\_ and \_\_\_\_\_, "Cooperation, Harrassment, and Involuntary Unemployment: An Insider-Outsider Approach," *American Economic Review*, March 1988, 78, 167–88.

# Fairness and Unemployment

By GEORGE A. AKERLOF AND JANET L. YELLEN\*

There have been natural reasons for the recent development and partial acceptance of efficiency wage theory. *Some* theory embodying payment of more than market-clearing wages is necessary to explain important features of the labor market, such as the existence of involuntary unemployment; the observed unwillingness of firms to reduce wages in the presence of job queues; the persistence of pay differentials for workers with seemingly identical characteristics; and the procyclical behavior of quits.

Alan Krueger and Lawrence Summers (1986a, b) and William Dickens and Lawrence Katz (1986a, b) have compiled impressive evidence in support of a central prediction of efficiency wage theory—that workers of identical characteristics receive different wages. These papers show the existence of large differences in wages across industries and occupations for workers after controlling for union status and observed worker and job characteristics. The impact of industry affiliation on wages is not just significant; it is also large, ranging in 1984 from a high of 38 percent above the mean for the petroleum industry to a low of 37 percent below the mean for private household services (Krueger-Summers, 1986a, p. 8). The leading neoclassical justifications for such differentials—unmeasured labor quality, compensating differentials, and transitory labor immobility—cannot easily account for the patterns of observed wage differentials.

As Lawrence Summers has observed, in its most general form, efficiency wage theory is virtually tautologous: it states that firms re-

frain from wage cutting because they don't perceive it to be in their interest. But *why* don't firms perceive wage cuts to be in their interest? The leading efficiency wage arguments are based on maximizing behavior; they stress the role of high wages both as a worker discipline device and also as a turnover reduction device. While efficiency wages per se are a useful concept, the maximizing models fail to explain four well-documented empirical regularities. They fail to predict wage compression (Robert Frank, 1984; Jean Grossman, 1983), the strong positive correlation between industry profits and industry wages (Dickens and Katz, 1986a; Krueger and Summers, 1986b), and the inverse correlation between unemployment and skill. Particularly striking is the unpredicted correlation of industry wage premia across occupations (Dickens-Katz, 1986b). *All* workers in better-paid industries tend to receive positive wage premia. That is, the wages of secretaries and engineers are highly correlated across industries. Ease of supervision and the magnitude of turnover costs might well be correlated across industries for a given occupation explaining, for example, why, say, skilled machinery operators receive positive wage premia in most industries. But there is no obvious reason why, say, secretaries, should be harder to supervise in the chemical industry where pay is high, than in the apparel industry where pay is low.

This paper presents an efficiency wage model based on fairness which provides natural explanations for the four regularities described above. According to this model, in industries where it is advantageous to pay some employees highly, it is considered *fair* to also pay other employees well. This theory is more in concert with basic worker motivation than standard maximizing models in which workers hedonistically maximize utility dependent only on their pay and their effort. Such hedonism seems too simplistic.

\*University of California, Berkeley, CA 94720. We are grateful to the Alfred P. Sloan Foundation and the National Science Foundation for generous financial support under grant No. SES86-005023 administered by the Institute of Business and Economic Research of the University of California, Berkeley. We also thank Daniel Kahneman for valuable conversations.

Well-paid workers often grumble and shirk while poorly paid workers are often satisfied and hard working. All textbooks on compensation consider it self-evident that the most important aspect of a compensation system is its accordance with workers' conceptions of equity. Workers who consider themselves fairly treated are likely to work hard, and workers who consider themselves unfairly treated are likely to shirk.

Daniel Kahneman, Jack Knetsch and Richard Thaler (1986) have demonstrated that popular conceptions of fair wages and prices often diverge substantially from the prices of goods and the wages of labor which would clear competitive product and labor markets. A hardware store that raises the price of snow shovels after a snowstorm would be considered to have acted unfairly. Similarly, in the labor market, a firm that has been making money and that reduces the wages of its existing labor when there is unemployment in the area will also be considered to have acted unfairly.

Kenneth Arrow has written that economic theory is concerned with "the forces of greed and aggressiveness...not the best, but the strongest motives of humanity" (1972, p. 90). But, in contrast to the marketplace, where traders have little personal contact, in the workplace, where personal contact is close, other emotions such as "concern for fairness," pejoratively called "jealousy," are also important. The poets serve as witness to the importance of both greed (or desire) and jealousy (or hate). Following Robert Frost:

Some say the world will end in fire,  
Some say in ice.  
From what I've tasted of desire  
I hold with those who favor fire.  
But if I had to perish twice—  
I think I know enough of hate  
To say that for destruction ice  
Is also great  
And would suffice.

### I. Fairness and Unemployment

In the most rudimentary model of efficiency wages due to Robert Solow (1979), output with one unit of capital is a function

of labor efficiency units, which are the product of man-hours and effort expended, so that  $q = f(e(w)l)$  where  $q$  is output;  $e$  is effort;  $w$  is the real wage; and  $l$  is man-hours. The optimal wage for the firm, that minimizes labor cost per efficiency unit, occurs at the wage  $w^*$ , where the effort function has unit elasticity with respect to the wage. With the wage at  $w^*$ , the firm will demand labor,  $l^*$ , up to the point where the marginal revenue product of labor equals the wage,  $w^*$ . In mathematics, this condition is  $e(w^*)f'(e(w^*)l^*) = w^*$ . If there are  $K$  units of capital and  $N$  units of labor inelastically supplied, unemployment will be  $N - Kl^*$ . The worker-discipline device and turnover models serve to derive the nature of the  $e(w^*)$  function from first principles rather than to posit it in an *ad hoc* way.

An alternative efficiency wage model that considers fairness could be formulated as follows. Suppose that there are two types of labor, call them  $l_1$  and  $l_2$ , and suppose also that output with one unit of capital can be expressed as:  $q = e(\cdot)f(l_1, l_2)$ . Let us follow Solow in trying on for size an *ad hoc* assumption about the effort function  $e(\cdot)$ , whose argument is as yet unspecified. Assume that  $e$  depends on the variance of wages paid by the representative firm:  $e = e(\sigma^2(w))$ . The simple rationale for this dependence of effort on wage variance is that firms with less variance in their compensation will have more harmonious labor relations and thus achieve higher output per worker. The firm's profit function is now:  $\pi = e(\sigma^2(w))f(l_1, l_2) - w_1l_1 - w_2l_2$ .

It is easy to see that in such an economy the higher paid labor will always receive a market-clearing wage and will have no involuntary unemployment. A reduction in the wage for higher-paid employees unambiguously benefits the firm. As the wage to higher-paid workers declines, effort increases as a consequence of the decrease in the variance of wages; in addition, the firm's wage bill declines. Without loss of generality, we may assume that  $l_1$  is the higher-paid labor type and that its wage clears the labor market. The firm will then choose  $l_1$  and  $w_1$  at market-clearing levels, and it will choose  $l_2$  and  $w_2$  to maximize profits. With



an interior solution to both of these maximization problems, it will choose:  $\partial\pi/\partial l_2 = 0$  and  $\partial\pi/\partial w_2 = 0$ . If these two conditions have an interior solution and  $N_2 > Kl_2^*$ , there will be unemployment of type 2 labor. The wage differential between types 1 and 2 labor will be "compressed" in comparison with the perfectly competitive equilibrium.

A very slight change in the model yields a result exactly analogous to Solow's. If one assumes that the effort of the lower-paid labor depends on the wage gap, but the effort of the higher-paid labor is independent of the gap, the firm then maximizes  $\pi = f(l_1, e(\sigma^2(w))l_2) - w_1l_1 - w_2l_2$ . Here  $w_1$  and  $l_1$  are at market-clearing levels. And the firm's optimal choice of effort and wage for type 2 labor satisfies the familiar Solow condition. With an interior-maximizing solution, the elasticity of effort with respect to the wage of type 2 labor (due to the dependence of effort on the variance of wages) will be unity.<sup>1</sup>

This model provides a straightforward rationale for wage compression and easily explains the observed positive correlation of industry wage premia across occupations. In industries where market forces compel firms to pay high wages to their engineers, considerations of equity will generate high pay to secretaries as well. The model also offers an explanation why unskilled workers have much higher unemployment rates than skilled workers. The model predicts that groups of workers with high pay will have lower unemployment than groups with low pay. Because high skill and high pay are typically correlated, this means that higher-skilled groups will usually have lower unemployment rates than lower-skilled groups. Occasionally, however, due to oversupply, highly skilled groups have high unemployment rates. Classical musicians are an example. Despite their skill, classical musicians are in considerable supply, and, as a result, their pay is low. They also have high rates of unemployment. Our model, which pre-

dicts that low-paid workers will experience high unemployment, is consistent with this anomaly.

The preceding model focuses on the adverse reaction of workers to inequitable pay disparities among employees. The same destructive forces of jealousy will also operate when there are "unfair" disparities between the earnings of the firm and the earnings of workers. Thus, the natural extension of our model, to include a relation between effort, firms' earnings, and workers' earnings, predicts the observed positive correlation between profits and wages. The correlation between profits and wages gives added reason for the positive correlation of industry wage premia across occupations.

## II. The Fair Wage/Effort Hypothesis

The Solow model is useful in characterizing the type of unemployment which occurs when the labor market is in equilibrium. Nevertheless, it is not a complete model because the relation  $e = e(w)$  is *ad hoc*. Similarly, the effort/wage-variance model predicts an unemployment equilibrium which accords with the stylized facts of the labor market discussed above. But this model, like the Solow model, lacks an adequate microfoundation for its effort function.

Social exchange theory in sociology and equity theory in psychology both provide a rationale for a precise relationship between effort and fairness. (For a similar rationale based on worker cohesion, see David Levine, 1987.) According to equity theory, in interpersonal exchange, the ratio of the perceived value of outcomes to inputs will tend to be equal. It is important to note that this equality is based on *subjective valuations* and not on *market valuations*.

In the context of a wage contract, equity theory suggests that the perceived value of an individual's labor input,  $e\hat{w}$ , should equal the perceived value of the compensation received,  $w$ , where  $w$  is the actual wage and  $\hat{w}$  is the wage which is considered fair compensation for "normal" effort. Assuming that workers supply "normal effort,"  $e = 1$ , if they are compensated at least fairly, and lower effort if they receive "unfair wages," equity

<sup>1</sup>This condition assumes that  $\sigma^2(w)$  is calculated with constant amounts of  $l_1$  and  $l_2$  labor. If  $\sigma^2(w)$  is calculated on the assumption of varying  $l_1$  and  $l_2$ , no simple formula obtains.

theory implies that  $e = \min(w/\hat{w}, 1)$ . Insofar as the actual wage  $w$  falls short of the fair wage  $\hat{w}$ , the worker will withdraw effort so as to maintain equity between the subjective value of input and the subjective return. We call the preceding behavior with respect to effort the *fair wage/effort hypothesis*. Under this hypothesis, if production depends on labor efficiency units and if firms have even the smallest preference for paying the fair wage (rather than less than fair wages), firms will always find it worthwhile to pay at least the fair wage  $\hat{w}$ .

There are many possible theories concerning the determination of  $\hat{w}$ ; but any theory which leads to an equilibrium value of  $\hat{w}$  for some class of labor greater than the market-clearing level will result in unemployment. The work previously mentioned of Kahneman et al. tells us that such discrepancies do occur; popular views of fair wages do not coincide with market clearing. A theory, for example, in which workers use the pay of higher-paid workers at the same firm as their anchor in determining fair pay, gives exactly the type of result obtained in the last section. Lower-paid workers are unemployed while higher-paid workers receive market-clearing wages.

The fair wage/effort hypothesis serves three useful purposes. First, it gives a natural functional form for the relation between effort and wages—while leaving considerable room for variations in the theory dependent on the determination of the fair wage. Second, it grounds the wage-effort function in theory from two social sciences, both reflecting the basic human desire to “*get even*.” Third, equity theory (and to a lesser extent social exchange theory) has been subject to empirical verification: for the psychologists through experiments, and for the sociologists through field studies.

A number of experiments have been performed to test the implications of equity theory for the inequity-performance relationship. Surveys of these experiments suggest that they are usually supportive of equity theory. In a typical study, subjects are asked to perform a task, such as proofreading, and offered an “hourly” or “piece” rate designed to induce a clear feeling of over- or under-

payment relative to a salient comparison group. The outcomes of these experiments typically accord with the predictions of equity theory: underpaid hourly subjects usually decrease their labor input to achieve an input-outcome balance. For example, Edward Lawler and Paul O’Gara (1967) compared the performance of workers who were paid the “going piece rate” vs. an underpaid rate for conducting interviews. They found the lower-paid workers produced work of substantially lower quality (but greater quantity) and had reduced self-esteem. Further evidence comes from a study by Robert Pritchard, Marvin Dunnette, and Dale O. Jorgenson (1972), who hired men to work for a manpower firm they set up. After the workers had been on the job for three days, the firm changed the pay with some subjects receiving increases and some subjects receiving decreases in pay. Those who got pay cuts were angry and performed less well in their work. All of these experiments suffer from two unavoidable difficulties: the anchor point of the fair wage must be inferred and the experiment by its nature is of short duration. Whether the observed behavior consistent with the fair wage/effort hypothesis will continue over long periods of time has not been inferred from experimental data.

Sociologists have explained field study observations with social exchange theory. The most famous of these is a study by Peter Blau (1955). In a federal bureaucracy agents varied in expertise. He found that agents of equal expertise frequently consulted one another, but only rarely did agents of average expertise consult the experts. Blau asked why the agents of average expertise did not consult the experts more frequently. The experts were observed to be friendly and always helpful, if asked. According to Blau’s answer, the two sides of the exchange had equal value—the average consulting agent receiving the advice of the expert with the expert receiving gratitude and respect. But such gratitude and respect had diminishing returns. Equilibrium occurred where the exchange was considered fair by both sides; the payment in gratitude and respect to the experts was of equal value with the aid given by the experts.

The scientific findings of social science notwithstanding, the key advantage of equity theory is its accordance with common sense. According to George Homans: "The more to a man's disadvantage the rule of distributive justice fails of realization, the more likely he is to display the emotional behavior we call anger" (1961, p. 75). Or, more prosaically, if people do not get what they think they deserve they get mad. This is the basic simple proposition underlying the fair wage/effort hypothesis. If  $w$  falls short of  $\hat{w}$ , people will be angry. As a result of this anger, their effective labor input is reduced below the level they would put in if fully satisfied.

### III. Theories of Fairness

The fair wage/effort hypothesis specifies the relation between effort and wages conditional on the fair wage  $\hat{w}$ . Different theories of the fair wage  $\hat{w}$ , will yield different theories of unemployment. Our longer paper (1987) constructs a model with two groups of workers, a high-paid group and a low-paid group. The low-paid group regard their fair wage as an average of the wages of the high-paid group in the same firm (who constitute part of their reference group) and the wages they would be paid if the market cleared. This formulation is a compromise between theories in which market forces completely determine fairness (in which case fair wages and market-clearing wages would exactly coincide) and theories in which sociological theories completely determine fairness (in which case the wages of some reference group would determine the fair wage). The micro model is surprisingly complex, but the macro intuition of the variance-efficiency wage theory of the last section survives. There is wage compression and the low-paid workers are those with unemployment.

The assumption that low-paid workers base their notions of fairness on the pay received by more highly skilled workers is undoubtedly a substantial simplification of how individuals realistically form conceptions of equity. Most workers in fact feel that fairness requires a relation between remuneration and performance. Workers with

low skill do not consider it fair to receive the identical wage as workers who are obviously more skilled. The theory of unemployment described here does not, however, rely on the simplistic assumption that workers view wage equality as fair. What is important is that workers regard a fair wage system as one with pay differentials which are more compressed than productivity differentials. An interesting study by Herbert Meyer (1975) suggests that this is indeed the case. He found in surveys that workers have a systematic tendency to view their own performance as superior. In four surveys he found that between 68 and 86 percent of workers viewed their performance in the top quartile.

### IV. Conclusion

Careful use of sociological theory, which thankfully also coincides with common sense, produces efficiency wage models of unemployment. The key to such models is the fair wage/effort hypothesis, that says that effort is proportional to the wage for workers paid less than the subjectively determined fair wage. The fair wage/effort hypothesis comes naturally from equity theory and social exchange theory. A simple hypothesis regarding the determination of the fair wage  $\hat{w}$  explains why unemployment is greater for low-paid labor groups (even of high objective skill). Efficiency wage theories based on fairness do not suffer from the criticism levelled at more neoclassical theories that if all agents are fully maximizing more complicated contracts will eliminate involuntary unemployment. They are also consistent with observed industry and occupational wage differentials.

### REFERENCES

- Akerlof, George and Yellen, Janet, "The Fair Wage/Effort Hypothesis and Unemployment," mimeo., University of California, Berkeley, July 1987.
- Arrow, Kenneth, "Models of Job Discrimination," in A. H. Pascal, ed., *Racial Discrimination in Economic Life*, Lexington: D. C. Heath, 1972, 83-102.
- Blau, Peter M., *The Dynamics of Bureaucracy*:

- A Study of Interpersonal Relations in Two Government Agencies*, Chicago: University of Chicago Press, 1955.
- Dickens, William and Katz, Lawrence, (1986a) "Interindustry Wage Differences and Industry Characteristics," NBER Working Paper No. 2014, September 1986.
- \_\_\_\_\_ and \_\_\_\_\_, (1986b) "Industry and Occupational Wage Patterns and Theories of Wage Determination," mimeo., March 1986.
- Frank, Robert, "Are Workers Paid their Marginal Products?," *American Economic Review*, September 1984, 74, 549-71.
- Grossman, Jean, "The Impact of the Minimum Wage on Other Wages," *Journal of Human Resources*, Summer 1983, 18, 359-78.
- Homans, George C., *Social Behavior: Its Elementary Forms*, New York: Harcourt Brace Jovanovich, 1961.
- Kahneman, Daniel, Knetsch, Jack and Thaler, Richard, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, September 1986, 76, 728-41.
- Krueger, Alan and Summers, Lawrence, (1986a) "Efficiency Wages and the Inter-Industry Wage Structure," mimeo., 1986.
- \_\_\_\_\_ and \_\_\_\_\_, (1986b) "Reflections on the Inter-Industry Wage Structure," Harvard Institute of Economic Research Discussion Paper No. 1522, July 1986.
- Lawler, Edward E. and O'Garra, Paul W., "The Effects of Inequity Produced by Underpayment on Work Output, Work Quality, and Attitudes Toward the Work," *Journal of Applied Psychology*, October 1967, 51, 403-10.
- Levine, David, "Cohesiveness and the Inefficiency of the Market Solution," mimeo., Harvard University, March 1987.
- Meyer, Herbert, "The Pay for Performance Dilemma," *Organizational Dynamics*, Winter 1975, 3, 39-50.
- Pritchard, Robert D., Dunnetter, Marvin D. and Jorgenson, Dale O., "Effects of Perceptions of Equity and Inequity on Worker Performance and Satisfaction," *Journal of Applied Psychology Monograph*, February 1972, 56, 75-94.
- Solow, Robert, "Another Possible Source of Wage Stickiness," *Journal of Macroeconomics*, Winter 1979, 1, 79-82.

# THE ECONOMICS OF THE ARMS RACE<sup>†</sup>

## Self-Interest and National Security

By WILLIAM G. SHEPHERD\*

The inner mechanism of global competition remains much as Adam Smith defined it two centuries ago: nations interact, while seeking to "win" by gaining higher wealth. Self-interest in gaining wealth drives the process, although wars and other deviations often occur. In this setting, economic concepts of competition, benefit-cost criteria and risk can help in defining efficient choices among policy tools (including military activities).

From research using those concepts (my study with Theodora Shepherd, 1986), I will discuss several tentative conclusions about the U.S.-USSR rivalry. 1) The global competitive process (let us call it Process 1) has properties which appear to make it stable. 2) The United States and USSR appear to possess inherent security from conquest by each other. 3) Each country's efficient limit on military spending can be analyzed as an analog of payments for insurance, to raise national security. In that context, inherent security reduces the U.S. and USSR's efficient levels of armaments. 4) Military entropy occurs; global wealth is subject to a general process (Process 2) which subtracts it into military forms and warfare.

### I. Process 1: Stable Competition and Zones of Interest

Nations expand their evolving sets of resources by making choices which weigh costs and benefits, broadly defined. Of course deviations are often caused by self-serving leaders, paranoia, ideology, and war fevers.

<sup>†</sup>*Discussants:* F. M. Scherer, Swarthmore College; Lloyd J. Dumas, University of Texas-Dallas; Walter Adams, Michigan State University.

\*Professor of Economics, University of Massachusetts, Amherst, MA 01002.

As a check on them, the criteria of wealth maximizing provide the populace an indispensable basis for defining and resisting deviance by their officials.

Each country also seeks some degree of influence on other countries so as to promote its ultimate commercial interests, by means of various foreign/military policies (Thomas Schelling, 1966). How far should it try to extend its influence, in light of the costs and benefits? Influence is costly to gain, and its marginal costs rise with distance. Conversely, the marginal benefits decline with distance.<sup>1</sup>

Each country therefore is surrounded by a finite *natural zone of influence*; its boundary lies where the marginal costs and benefits of influence are equal. Any nation's spending to extend its influence outside that natural zone will reduce the nation's wealth, because it exacts costs greater than the benefits, and increasingly so at greater distances.

This natural zone is probably small for most countries, generally lying in the nation's own region. If (as seems likely) the United States and USSR have limited natural zones in their regions, then most of the world lies outside both of them. U.S. or USSR foreign/military actions in that large neutral space would merely reduce their own wealth and be self-defeating in the wealth-seeking competition. Yet these nations (and many others) often do seek influence outside the natural zones: thus, the U.S. tries to "pro-

<sup>1</sup>Influence over distant countries is more costly because the ally's opportunity cost is higher compared to allying with a country in its own vicinity. As gravitation models of trade show, countries' trade levels vary inversely with distance. Variety among countries (in culture, language, and interests) causes distant alliances to be more expensive and difficult. The problems grow disproportionately with distance, because of simple plane geometry.

ject power" in far places with weapons surrounding the USSR, while the USSR pays subsidies to Cuba, etc. Such actions may be merely costly errors.

Generally, small zones and careful self-interest will promote stability in Process 1. Small U.S. and USSR zones would make their competition quite stable, much as the ancient Roman and Chinese empires co-existed neutrally. Efforts at "world domination" by either side could be ruinously costly. If populaces recognize these costs, they may control their leaders.

If the natural zones were large and overlapping, then direct conflict over "domination" would be chronic, and the competition could yield unstable, catastrophic outcomes. The diminishing net yields of influence, as distance increases, tend to prevent that.

## II. Inherent National Security, Based on Self-Interest

Within this general setting, the U.S. and USSR appear to be inherently secure from each other. The true risks are, instead, a purely military doomsday exchange, from error or lunacy, and the dissipation of wealth by military entropy in Process 2.

The U.S. and USSR's inherent security derives from self-interest, is of long standing, and is often tacitly acknowledged. But both countries' policies and rhetoric express the opposite: that the "opponent" is poised for invasion or nuclear blackmail, deterred only by military might.

To assess this, consider first the possible self-interest of the U.S. or USSR in subduing each other. The three alternatives are 1) to prevail, take control and draw large economic tribute from the victim (or at least eliminate it as an economic rival), 2) to prevent the other country from trying and/or achieving such a takeover, and 3) to avert an attack caused by error or madness. Reason 1 is critical; if it is known to be mutually invalid, then reason 2 loses relevance. Only reason 3 would remain to justify military protection.

Consider reason 1: there are three main ways in which the USSR might try to subdue and drain the U.S. (or vice versa). They are,

in ascending order of likely success: 1) *Invasion*. Warfare is followed by occupation, control of the economy, and a managed flow of tribute. 2) *Direct Control*. Under USSR nuclear blackmail, occupation is peaceful and largely civilian, with economic control and a managed flow of tribute. 3) *Pure Ransom*. Under nuclear blackmail, tribute is paid but with no occupation or direct control.

It is widely agreed that case 1 could not succeed. Damage to both sides would be too great, and little tribute could be wrung from a crippled economy. Cases 2 and 3 are at least conceivable, and they have been the ostensible rationale for both countries' armaments. But if cases 2 and 3 were sharply negative, then reasons 1 and 2 would lose relevance. For example, if costs were known to be five times the benefits, then an attack would gravely damage the populace's interest in maximizing wealth. Only by knowing this may the populace seek to restrain its leaders.

Elsewhere, some wars can pay, at least in the short run. In history, the common case is a quick raid on a neighbor to seize plunder and slaves; in classical Greece, such expeditions were organized explicitly as business ventures. But in a USSR move on the U.S. (by nuclear blackmail), the old-style plunder would be virtually nil; the USSR would mainly leave the U.S. resources *in situ*, to support U.S. production as the source of tribute.

Meanwhile the USSR would bear large direct costs, for weapons and control. Far larger still would be the many decades of high military spending as the USSR, a pariah nation, sought to avert or repel retaliation by the U.S. and other countries. Finally, the eventual retaliation (overtly or covertly) would cause disintegration and destruction in the USSR.

Tribute is therefore the central issue; could it conceivably be large enough to outweigh the costs and risks? Consider the main conditions which reduce the possible tribute obtained by country *A* from country *B*:

1. Distance (which makes control and the transport of goods difficult).
2. Large size of *B* (which requires large-scale, complicated controls).

3. Complexity of *B*'s economy (which makes controls ineffective).

4. Competitive market structures in *B* (low concentration makes difficult the control of the numerous enterprises).

5. Higher technology in *B* (which make controls weaker and self-defeating).

6. Involvement of *B* in foreign trade and ownership (which weakens the attempts to control, and draws in other countries).

7. Immobility of *B*'s resources (which prevents removal).

8. Widespread weapons and skills for popular resistance in *B*.

9. Possibilities for later reversal or revenge by *B* (which terminates control and imposes damage on *A*).

We have shown in some detail (myself and Shepherd) how these conditions would block U.S.-USSR tribute going in either direction. Distance and size alone would probably prevent significant tribute in real terms, but the other conditions also appear to be conclusive. The USSR's main source of possible gains would be our high-technology industries, but they would be precisely the ones most vulnerable to collapse, as innovation, investment and production quality fell. Civilian resistance (by withdrawal of effort, mass disobedience, and guerrilla violence) could cause endemic chaos (see Gene Sharp, 1985, among others).

One imagines baffled USSR officials on Wall Street and in boardrooms, attempting to monitor or control the investment process while high-technology industries shrivel, innovation ceases, workers' efforts decline by half or more, guerrillas disable much of the infrastructure, other countries intervene, and so on. A formal surrender by U.S. leaders could not prevent chaos at the grass roots, holding tribute to a trickle.

Self-interest therefore makes the ingrained human fear of conquest irrelevant for the U.S. and USSR. History confirms this economic analysis. Past conquests have yielded large tribute when the countries are neighbors, the victim is smaller, has less-complex low-technology industries which are concentrated, is insulated from world markets, has easily removable plunder, an unarmed and docile populace, and no way to retaliate.

Point by point, the U.S. and the USSR are both the opposite, particularly the U.S.

In this century, tribute has usually been small even in favorable cases. Large instances are only two: Germany's control of France during World War II, and the USSR's control of eastern Europe. France yielded Germany only minor benefits, though it was ruthlessly exploited (Alan Milward, 1970). Since about 1975, the USSR has probably endured net losses from its nearby satellites, not positive tribute (Michael Marrese and Jan Vanous, 1983; Charles Wolf et al., 1983). Neither case remotely approaches the difficulties faced by a U.S.-USSR attempt at conquest; even a USSR move to establish control over Western Europe would be folly.

Of course, inherent security does *not* guarantee against purely military attacks, by error, sinister leaders, or madness. Those arise from other causes, including the mere existence of the weapons, not from self-interest. In fact, our analysis shows certain classes of weapons to be the cause of danger rather than its cure.

### III. Efficient Levels of Military Spending

Each country's main objective is to produce, trade and accumulate efficiently, under a reasonable degree of security. Military spending may protect the populace and its wealth, but it drains that wealth into economically sterile military uses. How much weaponizing of the wealth is enough? The choice can be analogized to insurance choices by individuals (see J. D. Hey, 1979; Nicholas Rescher, 1983).<sup>2</sup> This approach also clarifies the savings provided by U.S.-USSR inherent security.

There are three main elements: the exposure to loss, *E*; the probability of loss, *p*; and the payment for loss avoidance or compensation, *I*. In the simple case, the maximum efficient payment is:  $I = p \cdot E$ . The payments may go either to loss prevention

<sup>2</sup>This is not the pure case of insurance against predictable risk. Military costs are partly for prevention, which preserves wealth. Yet, like locks and fire alarms, military measures are parallel to the restoration of wealth after a loss.

(for example, locks for a house) or insurance (to repay losses), or both.

In *very* brief summary, the efficient level of coverage is a marginal choice, aligning the cost of protection with the perceived scale of loss. Assuming convexity (declining marginal objective tradeoffs between spending and security), the specific optimum level can be formally defined for simple cases, although risk attitudes vary among individuals and may be unknowable.

A nation's exposure  $E$  includes the destruction (of physical and human assets; capitalized net losses of production and trade; etc.) expected under an attack from outside. Another important loss is of assets (land, equipment, slaves, etc.) taken by the winner, plus any tribute. The literature rarely mentions such tribute, but it would have to be the main incentive for any attack based on real interests (as Section II noted).

By influencing the populace's beliefs about danger (the  $p$  and  $E$  values), leaders shape the perceived need for armaments. An overstatement of those risks will raise  $I$ , the wealth that is weaponized. The weapons costs will understate the true cost, if 1) the weapons themselves raise  $p$ , by stirring fear and greater escalation, 2) war fever sacrifices social values, or 3) widespread weapons production reduces industrial efficiency.

In situations involving a number of small countries, the efficient arms level for each one may be indeterminate over a wide range, even under full information. Moreover, deep grievances among close neighbors (as in the Middle East) often enhance both  $p$  and  $E$ , and therefore  $I$  (Gavin Kennedy, 1983).

In the U.S.-USSR case, *inherent security detaches this insurance logic*. If successful conquest is impossible both ways, then both superpowers'  $p$  and  $E$  values (and therefore  $I$  values) approach zero. Protective weapons (such as the U.S. triad of bombers and land-based and submarine-based missiles, plus much of the Navy) could be omitted without reducing U.S. security.<sup>3</sup> The U.S.

military resources might have other purposes (though usually not positive net yields outside the natural zone of interest: recall Section II), but national security would not be one of them.

Inherent security also calls for a reassessment of the meaning of "the national interest." Policy gains abroad are often just small commercial advantages (a few dollars off the price of oil, some airport slots in Europe, some more computer sales in Japan, etc.). Each one should meet cost-benefit criteria as stringent as those applied to domestic programs. In current debates, they are almost entirely free of cost-benefit tests.

#### IV. Process 2: The Erosion of Wealth

Evidently, a probing of self-interest and wealth can clarify national security, showing some of the sources and effects of excess weaponizing. Not only may U.S. and USSR strategic weapons be functionless, but the weapons themselves create risks, by raising the probability and scale of possible purely military attacks.

Parallel to the benign Process 1, a negative entropic Process 2 appears to function. It tends to dissipate wealth by military spending, by creating risks which deter investment and distort allocation, and by actual destruction in warfare. Process 2 operates in four main directions.

1) *Military Defense*. A wealthy country is the natural prey of others, and so it will convert some of its wealth into military protection, as noted in Section III. Moreover, jingoistic leaders (seeking to create popular loyalty against supposed alien threats) will overstate the risks, leading to excess weaponizing of the wealth. Also, weapons-selling countries will encourage their buyers to over-acquire weapons.

2) *Allies*. As has always occurred in statecraft, "allies" impoverish the alliance leader. They bargain for payments, trading

<sup>3</sup> Moreover, strategic weapons may already be made irrelevant by micro-military devices, such as suitcase nuclear bombs and briefcase chemical weapons. Rela-

tively few of these, planted in each others' major cities, could supersede the arrays of bombers and rockets, and at a trivial cost. They would leave the proposed SDI system with no weapons to deter.



privileges and gifts in return for "influence" (which is often barren, as Section I noted). They encourage a disbelief that inherent security exists, so as to enlarge their drainage of the super-powers. A few such allies (in Western Europe, the Philippines, Turkey, Japan, etc.) may deplete even a large store of national wealth and cause chronic economic retardation.

3) *Induced Weaponizing and Wars.* The costs are magnified when rivals succeed in inducing each other into extra weaponizing, payments to allies, and irrelevant wars. Increasing such self-injury is traditionally a major element of statecraft. In this looking-glass world, for example, the USSR may "win" in Vietnam, by inducing functionless waste and destruction by U.S. officials. Nicaragua, Cuba and Afghanistan are other self-harming examples, all probably stimulated by the other side's manipulative actions. And with SDI, the U.S. may successfully induce the USSR into functionless wastes even larger than our own; unless, instead, the USSR understand SDI's irrelevance and are merely inveigling U.S. officials into burdening the U.S. economy with it.

4) *Domestic Decay.* Wasteful actions abroad may stir domestic protests, which reduce social and economic stability. Though they may raise patriotism, rattling the saber and actual warfare can disturb society and cause economic deterioration.

Process 2 appears to function widely, in spending, warfare and side effects. In some nations it fully dissipates the wealth created by Process 1. The concepts outlined here may be useful in defining the costs and showing how to avoid them.

#### V. Further Research

Each of these concepts and tentative conclusions needs further research. The costs and benefits of influence, and the size of natural zones of influence, are largely untested. The likely gains and costs of attempted U.S.-USSR conquest could be esti-

mated, under alternative situations, to indicate the degree of actual inherent security. The factors affecting tribute need study, and historical cases of conquest may offer rich lessons.

The uses of insurance logic can be explored, and the residual interests (other than security) need to be clarified. Finally, Process 2's erosion of wealth has many dimensions needing study. Is it like Malthus's Law, endemic and universal even where it does not currently prevail *in toto*?

Further study seems promising, one might even say obligatory. The economist's simple cost-benefit comparisons can illuminate large areas of dark, ideological, geopolitical pseudoreality.

#### REFERENCES

- Hey, J. D., *Uncertainty in Microeconomics*, New York: New York University Press, 1979.
- Kennedy, Gavin, *Defense Economics*, New York: St. Martin's Press, 1983.
- Marrese, Michael and Vanous, Jan, *Soviet Subsidization of Trade with Eastern Europe: A Soviet Perspective*, Research Series No. 52, Berkeley: UC Institute of International Studies, 1983.
- Milward, Alan S., *The New Order and the French Economy*, Oxford: Clarendon Press, 1970.
- Rescher, Nicholas, *Risk*, New York: University Press of America, 1983.
- Schelling, Thomas, *Arms and Influence*, Cambridge: Harvard University Press, 1966.
- Sharp, Gene, *Making Europe Unconquerable: The Potential of Civilian-Based Deterrence and Defense*, Cambridge: Ballinger, 1985.
- Shepherd, William G. and Shepherd, Theodora B., *The Ultimate Deterrent: Foundations of US-USSR Security Under Stable Competition*, New York: Praeger, 1986.
- Wolf, Charles, Jr. et al., *The Costs of the Soviet Empire*, R-3072/1-NA, Santa Monica: Rand, 1983.

# Economic Consequences of the Arms Race: The Second-Rate Economy

By SEYMOUR MELMAN\*

The United States has been transformed into a second-rate industrial economy. While a depletion process in U.S. industry was identified as early as 1965 (see my 1965 study), the full quality of that process took a while to unfold. In 1987, it is useful to identify the following characteristics of a first-rate industrial economy against which the depletion process can be gauged: 1) the ability of the industrial system to offset cost increases of every sort by productivity growth; 2) the ability to pay high and rising wages while producing marketable goods; 3) vigorous research in basic science and in the technologies; 4) the availability of an increasingly competent production support base (infrastructure); 5) having the use of a currency of stable, meaning predictable, value; 6) having the capability for organizing people for productive work 7) as a result, enjoying a rising level of living.

From 1915 to 1950, U.S. industrial firms enjoyed sufficient productivity increases to offset a fivefold hourly earning increase to industrial workers, while only doubling the prices of all "metals and metal products" (see my 1956 study, p. 152). During the same period, it was characteristic that the prices of machines, notably machine tools, increased less rapidly than did the wages to industrial workers. That pattern made the purchase of new machine tools attractive on a continuing basis to machine users. Accordingly, American industry was known worldwide as well equipped with high-performance, high-productivity machine tools.

During a century of U.S. industrialism, 1865-1975, U.S. industry paid the highest wages in the world to industrial workers. The level of productivity was sufficient to

offset the wage while the firms held domestic markets and were competent suppliers abroad as well (see my 1983 book, ch. 10). At the same time there was expanding basic science research in the United States, and a notable interest in the technologies such that the machinery-producing industries and their clients were well-served with a considerable flow of new ideas.

The infrastructure in the United States was steadily improved. The railroad system, water supply, road networks, means of communication (telegraph, telephone, postal system), power supply, the education system—all served as a base of support for a widening and an increasingly productive industrial system. The U.S. dollar had a relatively stable, predictable value. It should be recalled that it was only relatively recently that the gold backing and convertibility of the dollar to gold were removed.

American industrial managements were also aggressively competent in organizing people for productive work with technologies that were shaped to suit managerial criteria. The idea of the assembly line, "Fordism," Frederick W. Taylor's "scientific management," national industrial unionism—all played a part in marshaling and operating the largest industrial labor force in the world.

The average level of living in the United States compared favorably with that of Europe, the main origin of immigrants to America, and also showed a central tendency of steady growth despite the fluctuations of the "business cycle."

By the 1980's, these conditions have been checkmated. The core of these seven characteristics is clearly the pattern of productivity growth. It was the productivity growth that made possible the cost offsetting, the high wage rate, the competitive price, and finally, the rising level of living. In the pres-

\*Professor Emeritus of Industrial Engineering, Columbia University, 305 Mudd Bldg., New York, NY 10027.

ence of other necessary conditions, including vigorous *R&D*, a competent infrastructure, a stable currency—it was the growth in productivity that was the driving factor in this set of conditions that constituted a first-rate industrial economy.

What happened to productivity growth? The average percent increase per annum in output per production worker in the manufacturing industries changed as follows: 1950's, 4.5; 1960's, 3.7; 1970's, 0.9; 1980–84, 2.3. During the 20 years before 1970, manufacturing productivity grew at an annual average rate of 4.1 percent—nearly three times as high as the 1.4 percent rate after 1970 (Lloyd Dumas, 1987). The ability to pay high and rising wages terminated by 1975. That was the last year in which the U.S. industrial wage was the highest in the world. By 1980, seven European countries were paying higher wages than U.S. industry (myself, 1983, p. 309; 1987).

The number of engineers and scientists per 10,000 of the labor force in 1965 was 64.1 in the United States; 24.6 in Japan; and 22.6 in West Germany. For 1977 I estimated the number of engineers and scientists *in civilian activity* per 10,000 in the labor force: U.S., 38; West Germany, 40; Japan, 50 (myself, 1983, pp. 170, 171; 1987). The United States had a larger gross number of engineers and scientists. But the intensity of their use on behalf of the civilian economy was substantially less than in the case of Japan, somewhat less than in West Germany. In 1970 (last year of available data) America's military-serving manufacturing industries employed an average of 7.4 scientists and engineers in research and development per hundred production workers. In civilian-serving manufacturing, the percentage was 1 percent (myself, 1983, p. 89; 1987).

The machine tool and the electronics industries have been important areas where the preemption of technical talent by the military has played a part in so weakening competitive position, as to hasten the decline, and even disappearance of major sections of those industries (Anthony DiFilippo, 1986; myself, 1983; 1987, Prologue).

The United States now lacks a modern rail system and a modern highway system in good repair (Pat Choate and Susan Walter, 1981). Many city streets are poorly paved. Between a fifth to a third of the highway bridges in the United States are rated as needing major repair. Decent housing is no longer available for millions. There is a growth of homelessness and hunger (Physician Task Force on Hunger in America, 1987). Important parts of the population draw water from aquifers that are contaminated. The national parks are in poor repair. The libraries are poorly operated. Waste disposal systems violate modern technical standards. The public school buildings of New York City require an expenditure of \$8 billion for decent repair.

With a currency of falling and unpredictable value, industrial planning is endowed with special hazards—apart from the risks of any new undertaking. International changes in the dollar's value have been combined with a deferred inflation process at home.

As the federal government sought to finance its enlarged military budgets by borrowing heavily after 1980, it set in motion a mechanism that effectively transferred inflation to the master commodity, the dollar. While the U.S. Treasury borrowed on a large scale, foreigners increasingly became lenders to the U.S. government. They bought U.S. dollars with their currencies in such large amounts that from 1980 to 1985, the price of the dollar rose 75 percent in relation to the average price of fifteen principal currencies. Hence the so-called "strong dollar" really meant a price-inflated dollar.

Two major effects were produced. By 1985, if U.S. producers wished to be as competitive (in the United States) vs. foreign producers of comparable goods—as they had been in 1980—they needed to reduce their U.S. costs and prices by as much as 57 percent. Since that could not be done, the United States suffered an epidemic of factory closings, which were hailed as a sign of a growing "service economy."

A further chapter in government borrowing to finance the Pentagon-induced budget

deficit is now unfolding. The price of the "strong dollar" is being collapsed by inter-governmental arrangement to forestall a precipitous drop, and by market forces impelled by the growing U.S. budget and trade deficits. The purchasing power of the dollar in relation to other currencies is then diminished, and the dollar prices of imports must rise. But there can be no corresponding growth in U.S.-based production because by 1985 many factories had been dismantled and their workers dispersed. With higher prices of imports, the purchasing power and average level of living of Americans will fall. In that way, the deferred and concealed inflationary effect of the government's financing of the military economy will be finally expressed (myself, 1987, p. 11).

The ability to organize people for work has diminished as 6 percent unemployment is treated as a success story.

For all wage and salary earners, the percent increase in average real income was: 1950's, 3.1; 1960's, 2.5; 1970's, 0.9; 1980-84, 0.0. A continuation of this trend portends a declining level of living, hitherto unheard of in American experience (Dumas).

Among the many factors that surely affect productivity growth, two elements dominate the scene. One is the presence of an ample supply of capital at an attractive rate. The second is a pattern of cost minimizing among the managers of industrial firms.

The largest single capital fund in the United States since 1951 has been the annual budget of the Defense Department. Economists ordinarily do not diagnose the characteristic of a modern military budget as a capital fund, but operationally, it is. By capital, I mean what we ordinarily understand as fixed or working capital in the industrial enterprise. The fixed capital is the money value of land, buildings, machinery; working capital is the money value of all the other resources that have to be brought to bear to make the enterprise function as a producing entity.

A modern military budget, when utilized, sets in motion precisely those resources. Hence, it is significant that the Comptroller of the Department of Defense informs us

that, cast in dollars of 1982 value, the DoD used \$7,620 billion of capital resources from 1947 to 1987 (U.S. Department of Defense, 1986, p. 125). The data on national wealth of the United States include a statistic for 1982 on the value of the nation's fixed, reproducible, tangible wealth. Excluding the money value of military material and household durables, that wealth statistic is a good approximation to the money value of the nation's producers' plant and equipment, and infrastructure. That money value in 1982 was \$7,292 billion (U.S. Department of Commerce, 1985, p. 461). The United States has used up, on military account (1947-87), a quantity of capital resources sufficient to renew the largest part of the industrial and infrastructure plant of the United States.

The second elemental requirement for a desirable rate of productivity growth is pervasive cost minimizing by management. If and only if decision making is oriented to improving efficiency in the use of resources, is there a real spur for productivity improvement. Cost maximizing, by contrast, is the characteristic condition of operation in the military serving industries of the United States where 35,000 establishments and over 100,000 subcontracting firms serve the Pentagon. Those managements have become indifferent to cost and operate by methods that ordinarily escalate cost and therefore price (see my studies: 1985, chs. 2, 3; 1983, ch. 5; 1987).

The consequence of this alteration in the microeconomic process that has spurred productivity growth has been a major loss to the United States of production capacity and competence. In 1979-80, 17 percent of the automobiles purchased in the United States were imports. By 1986, that was about one-third. Machine tool imports represented 25 percent of U.S. market sales in 1979-80, and 50 percent by 1987. Shoes: 45 percent were imported 1979-80; 86 percent in 1987. The prudent understanding is that there have been comparable degrees of loss of productive employment (myself, 1983; 1987, p. 200).

This collapse of production competence has included not only smokestack, but also high-tech industries—those that depend on

a strong R&D input. The high-tech group showed a favorable trade balance of \$27 billion in 1980. By 1986, they were recording a trade deficit (U.S. Congress, Joint Economic Committee, 1986).

For the first time in 200 years, competitiveness in American industry is not a micro problem of single firms. Incompetence in production has engulfed entire U.S. industries. By this yardstick Japan and Germany have won the Cold War, and the United States and USSR are the losers. As a consequence of its normal operations, that preempted capital resources and installed a cost-maximizing micro economy in U.S. industry, the federal government and Department of Defense have spearheaded the creation of a second-rate industrial economy.

None of this is to say that the tangle of U.S. economic problems that signal the decay of first-rate industrial status is the product of a single-cause system. Other factors surely play their part: managerial short termism; the idea that we are a "service" economy, a "postindustrial" society in which production is, by definition, unimportant; a naive belief in an inherent American technological superiority; belief that speculative profit represents real wealth; large and growing administrative costs, unrelated to efficiency in production; a consensus that favors military spending as a way to regulate the economy; etc.

It is to say that massive, sustained military spending is, qualitatively, the single most critical factor in the cumulative depletion of the industrial economy. If it is dealt with decisively, then the rest can be addressed. If that factor is unattended, then the rest is rendered unmanageable, and a process of continued decline is locked in place.

To illustrate: *The New York Times* of December 8, 1987 reported that U.S. intercontinental delivery vehicles—missiles, aircraft, submarines—are able to deliver 11,786 strategic warheads to the entire Soviet's 220 urban industrial centers. Hence, U.S. forces have more than 50 times overkill capability by this simple reckoning. A 75 percent reduction in the 1988 budget for operating and adding to this overkill capability would entail a budget saving of \$54.6 billions (Center

for Defense Information estimate) leaving an overkill capacity of 12, not less than an overkill of 50 times.

A new political economic factor has appeared owing to the depletion of both U.S. and Soviet economies by long-enduring military priorities. For the first time since World War II, a part of the ruling elites of both societies judge that in order to cope with domestic problems it is necessary to make substantial reductions in military budgets. Negotiation of large reductions in military budgets and armed forces are now conceivable and perhaps feasible.

Such a reversal would require a capability for eventually converting as much as \$295 billion in annual U.S. activity from military to civilian purposes. The necessary planning for economic conversion is defined by a 1987 bill sponsored by Representative Ted Weiss, Democrat of New York, and 50 more House members (U.S. Congress, House of Representatives, 1987).

It would set in motion a decentralized process emphasizing local responsibility and authority for using the people and facilities of factories, laboratories, and bases that now serve the military for civilian work instead. City, state, and federal governments would be marshaled to plan the capital investments needed to upgrade the decayed infrastructure. This would portend large new markets and work opportunities.

What is the prognosis if the military economy continues to dominate in the use of capital resources, and to spur cost maximizing? The second-rate economy will then become progressively less manageable and the conditions of a third-rate economy will evolve. Dumas has termed such a process, "undevelopment" (pp. 129–30), and John E. Ullmann defines a new group of economies; "fifth world" (1985, ch. 2). Their common feature is a lack of resources needed for repairing even key industries, and suffering a declining level of living. That condition is now found in industries ranging from trolley cars to consumer electronics to shoes. As the decay spreads, then, as in an unindustrialized country, teams of workers, technicians, and managers would have to be sent abroad to acquire needed skills, or foreign staffs

imported to the United States to train the natives.

### REFERENCES

- Choate, Pat and Walter, Susan, *America in Ruins*, Washington: Council of State Planning Agencies, 1981.
- DiFilippo, Anthony, *Military Spending and Industrial Decline, A Study of the American Machine Tool Industry*, Westport: Greenwood Press, 1986.
- Dumas, Lloyd J., *The Over-Burdened Economy*, Los Angeles: University of California Press, 1987.
- Melman, Seymour, *Dynamic Factors in Industrial Productivity*, Oxford, New York: Basil Blackwell, Wiley & Sons, 1956.
- \_\_\_\_\_, *Our Depleted Society*, New York: Holt, Rhinehart and Winston, 1965.
- \_\_\_\_\_, *Profits Without Production*, New York: Alfred A. Knopf, 1983 (Philadelphia: University of Pennsylvania Press, 1987).
- \_\_\_\_\_, *The Permanent War Economy*, New York: Simon and Schuster, 1985.
- \_\_\_\_\_, *Conversion from Military to Civilian Economy, An Economic Alternative to the Arms Race*, Washington: National SANE Education Fund, 1987.
- Ullmann, John E., *The Prospects of American Industrial Recovery*, Westport: Quorum Books, 1985.
- Physician Task Force on Hunger in America, *Hunger Reaches Blue Collar America*, Cambridge: Harvard School of Public Health, 1987.
- U.S. Congress, House of Representatives, H.R. 813: "A Bill to Facilitate the Economic Adjustment of Communities, Industries, and Workers to Reductions or Realignments in Defense or Aerospace Contracts, Military Facilities, and Arms Export, and for Other Purposes," 100th Congress, 1st sess., January 28, 1987.
- \_\_\_\_\_, Joint Economic Committee, *The U.S. Trade Position in High Technology: 1980-86*, Washington: Quick, Finan, and Associates, October 1986.
- U.S. Department of Commerce, Bureau of Census, *Statistical Abstract of the U.S. for 1985*, Washington 1985.
- U.S. Department of Defense, Office of the Assistant Secretary of Defense (Comptroller), *National Defense Budget Estimates for FY 1987*, May 1986.

# U.S. Military Power, the Terms of Trade, and the Profit Rate

By TOM RIDDELL\*

Most economists view military spending as the classic public good: the state assumes responsibility for the defense of the nation, its territory and its interests; and the resources for national security are allocated by political institutions. There is also a broad literature on the economic effects of military expenditures. Seymour Melman (1974) and Lloyd Dumas (1983), for example, argue that military spending has adverse economic impacts, including spreading inefficiency, reducing civilian *R&D* and investment, and contributing to inflation. But the Congressional Budget Office (1983) and Gordon Adams and David Gold (1987) have concluded that the negative consequences of military spending are exaggerated.

There is less attention, however, to the question of economic stimulants to military priorities. Adam Smith suggested the fundamental importance to capitalism of the state's provision of national defense, yet many modern economists are reluctant to admit the economic motivations for military spending. In the post-World War II period, with defense budgets ranging from 5 to 14 percent of GNP, this question may be too important to ignore. There may be important economic impediments to reducing military spending.

## I. Economic Factors Promoting Military Spending

Scholars have offered a wide variety of hypotheses to explain U.S. military spending. They focus on the composite domestic

and foreign factors (political, economic, technological and ideological) that stimulate military priorities. My 1982 paper identified a number of specifically economic explanations for continued high levels of U.S. military spending. Michael Reich (1978) points to the narrow economic interests of the military-industrial complex that continually demand increased military budgets. Paul Baran and Paul Sweezy (1966) argue that military spending prevents the economy from experiencing stagnation, and James Cypher (1974) adds that it has been used as an anticyclical tool in the postwar period. With a more international focus, R. P. Smith (1977) hypothesizes that military spending supports U.S. worldwide hegemony and consequently contributes to successful capital accumulation. A variant of this argument focuses on the Soviet threat to the international economic and strategic interests of the United States and the necessity of military might as a response.

In this paper, I propose an internationally based model of military spending. The argument is that international economic activity is important to U.S. capital accumulation and that U.S. military spending and power have supported the international dimensions of a postwar social structure of accumulation.

## II. The Connection Between International Economic Objectives and Military Spending

Samuel Bowles, David Gordon, and Thomas Weisskopf (1986) have argued the importance of a social structure of accumulation to the long-run growth process of capitalism. Capitalism is based on and driven by the capital accumulation process. Capital accumulation, stimulated domestically and internationally by competition for profits, requires continued expansion in the physical and social forces of the capitalist mode of

\*Associate Professor of Economics, Smith College, Northampton, MA 01063. I appreciate the efforts of numerous colleagues to shape my understanding of the economics of military spending in the post-World War II period. This work has been supported by the Jean Picker Fellowship and the Smith College Committee on Faculty Compensation and Development.

production. The process works best when political, economic, legislative, governmental, and international institutions complement each other to create a social structure of accumulation that encourages an environment supportive of the driving force of capital accumulation—profits. When profits can be generated in productive activity and realized in the sale of commodities on markets, capital accumulation is most likely to proceed and to produce economic expansion.

The postwar prosperity of the U.S. economy, for example, can be interpreted to be founded on a conjunction of favorable historical and institutional factors that promoted capital accumulation. If these factors deteriorate, capital accumulation slows down. Bowles et al. have tested components of this hypothesis. One of these factors was the position of the United States after World War II as the world's leading political, economic, and military power. It took the lead in creating an international financial and economic system intended to promote stable and growing trade. Generating profits through access to cheap raw materials and labor, overseas investments, and assured markets for goods and services was viewed as an important requirement of the capital accumulation process. An important support for this system was U.S. military superiority and power. The success of the postwar social structure of accumulation depended in part on U.S. military spending to create a worldwide "peacetime" military establishment.

While it is obvious that some firms and some industries in the United States have direct interests in stimulating international economic activity, there is a structural imperative as well. The system requires expansion of aggregate demand and markets for capital accumulation, and an open international economic system promotes that possibility. In addition, the benefits from such activities can be larger for the country (and the capitalist class of that country) that dominates the system. A military policy of extended deterrence, including both nuclear and conventional armaments, is used to

maintain dominance in the Western alliance, in the rivalry with the Soviet Union, and in the rest of the world in support of a stable international economic order.

In a sense, this is an extension of Charles Kindleberger's (1973) analysis of the international causes of the Great Depression. The inability of Great Britain and the unwillingness of the United States to assume leadership in stabilizing international finance and trade in the late 1920's and early 1930's was a key contributor to the length and depth of the Great Depression. As Fred Block (1977) has demonstrated, the lesson was learned; and a fundamental goal of Western economic policy in the aftermath of World War II was a stable and growing international political economy. This objective required a leader, and the United States fit the bill. In addition, leadership, as it had for Great Britain in a previous era, required military power through consistent military spending to create and maintain a global military establishment.

What support for this viewpoint is there? First, it is consistent with the interpretations of informed observers and policymakers. Second, it is supported by data on the importance of international activity to the U.S. economy. Third, a model can be specified that attempts to test some of its key propositions.

Diplomatic historians and U.S. policymakers have recognized the interconnections among international economic objectives, foreign policy and military spending. William Becker and Samuel Wells in assessing these linkages have concluded:

[T]he United States dominated the world economy for the first two decades after World War II. In part, this extraordinary influence flowed from America's predominant economic and military power within the international system. In a world struggling to rebuild after the devastation of the war, an unscathed America could expand its influence and spread its ideas and institutions through the non-Communist world essentially without challenge. But the multilateral economic institutions,



which the United States inspired and directed, also helped maintain American power by providing reliable foreign sources of raw materials and outlets for American exports and investments. [1984, p. 462]

My 1985 review of the public messages of Reagan Administration officials made it clear that international economic objectives are important determinants of foreign and military policy. President Reagan throughout his term has stressed a global vision of economic interests and the importance of U.S. military power. As he explained in a speech at Georgetown University's Center for Strategic and International Studies:

We began with renewed realism—a clear-eyed understanding of the world we live in and of our inescapable global responsibilities. Our industries depend on the importation of energy and minerals from distant lands. Our prosperity requires a sound international financial system and free and open trading markets. And our security is inseparable from the security of our friends and neighbors....

Gone are the days when the United States was perceived as a rudderless superpower, a helpless hostage to world events. American leadership is back. Peace through strength is not a slogan, it's a fact of life. And we will not return to the days of handwringing, defeatism, decline, and despair.

[1984, pp. 1-2]

In the postwar period, U.S. economic interests have increasingly developed international avenues for the expansion and exploitation necessary for the generation of profits and capital accumulation. Direct investment abroad has increased to more than \$225 billion; the ratio of imports to GNP is over 10 percent, while exports to GNP is a bit lower; and foreign profits have increased to more than a quarter of total corporate profits. Secretary of State George Schultz (1983) has emphasized the importance of international factors to the U.S. economy: 20 percent of U.S. jobs depend on trade, 40 percent of agricultural output is exported, 30 percent of U.S. exports go to the non-oil-

producing developing nations, 25 percent of U.S. imports come from the non-OPEC developing countries, the Third World supplies many important raw materials to U.S. industry; and U.S. multinational banks and the United States itself are fundamental components in the international financial system.

### III. An Internationally Based Model of Military Spending

Specifying a model that tests these propositions about the importance of the international economic system to capital accumulation in the United States and the supportive role of military spending is not an easy task, since the argument is largely qualitative and the significance of the military function may be basic and structural rather than temporal and instrumental. The importance of military power may rest with some minimum level of military presence around the world—requiring some constant amount of real military spending—and not depend on annual increments of military spending. Nevertheless, it is worth the effort; and there are some models that provide some guidance.

Most models use some measurement of military spending as the dependent variable and then examine the ability of different domestic and international factors to explain the variance in military spending over time or across countries. I propose a model that takes a different approach and examines the functional role of military spending and military power in the postwar social structure of accumulation (see my 1986 paper). I follow the example of Bowles et al. who focus on the profit rate as the primary indicator of the overall operation of capital accumulation and the relative success or failure of the social structure of accumulation. The theory posits that U.S. capitalism produces higher rates of profit when economic conditions are favorable and when various institutional factors favor U.S. capital at home and abroad. The power of capital over labor, U.S. citizens, and the rest of the world enhances the ability of the system to produce profits.

My model of the profit rate suggests that it is a residual that varies with the costs of production and capital's ability to exploit resources at home and abroad. I have estimated a model for the profit rate with annual time-series data for the United States from 1948 to 1981. The model takes the after-tax profit rate, a weighted average of the domestic profit rate for the nonfinancial corporate business sector and an estimated foreign profit rate, as the dependent variable. As independent variables to explain some of the variation in the profit rate, I use *CUMF* = the capacity utilization rate for manufacturing,  $\Delta ULC$  = the annual rate of change in unit labor costs (reflecting the combined effects of wage and productivity growth),  $\Delta RMP$  = the annual rate of change in the relative prices of crude materials, *TOT* = the terms of trade, and *MP* = an index of military power.

Theoretically, we would expect there to be a positive relationship between the level of activity in the economy, measured by *CUMF*, and the profit rate. Similarly, we would expect the profit rate, in the short run, to decline as wages grew faster than productivity and corporations fail to adjust prices immediately (producing a negative relation between the profit rate and  $\Delta ULC$ ). We would also expect a negative coefficient for  $\Delta RMP$ .

The ratio of the relative prices of exports to the relative prices of imports (*TOT*) is an independent variable that reflects the relationship of international economic factors to the profit rate. It rose through most of the immediate postwar period to a peak in 1968 and decreased almost continuously from then until 1980. When the relative prices of U.S. exports are greater than the costs of imports, there are economic advantages to the United States in the form of reduced relative costs of goods and services purchased from abroad. As Gerald Meier has suggested: "When a country's commodity terms of trade improve, its real income rises faster than output, since the purchasing power of a unit of its exports rises" (1982, p. 50). The better the terms of trade, the larger the gains from trade. Advantageous terms of trade facilitate capital accumulation; an improvement in the

terms of trade will tend to increase the profit rate. *TOT* is intended to capture movements in the political economic power of U.S. capital in the postwar period—reflecting the variation in the dominance of the United States in the international financial and economic order.

The variable *MP* is intended to capture the role of military power in U.S. international policy. It measures the number of incidents annually from 1948 to 1981 where the United States used its military forces without active conflict. Barry Blechman and Stephen Kaplan define these incidents: "...when physical actions are taken by one or more components of the uniformed military services as part of a deliberate attempt by the national authorities to influence, or to be prepared to influence, specific behavior of individuals in another nation without engaging in a continuing contest of violence" (1978, p. 12). To smooth out the pattern over time, a three-year backward moving average was calculated. *MP* was then lagged one year, on the assumption that the effects of the use of power would follow.

Although not a very complex measurement of U.S. military power, *MP* does reflect the outcomes of complicated balances, factors, considerations and relationships involved in foreign and military policymaking. Where the United States actually decided to use its worldwide military forces, it did so with purpose. My interpretation is that these incidents are an indicator of power, not weakness. Obviously, it is difficult to quantify the effective use of military power. It could be argued, for example, that if power is unchallenged, because it is effective, then forces needn't be used at all. On the other hand, the United States has consistently deployed its conventional and nuclear forces to achieve foreign and military policy goals throughout the post-World War II period. Consequently, I take an index of the use of forces short of violence as an indicator of the use of U.S. power when it was expected to produce results (both in the short run and the long run).<sup>1</sup>

<sup>1</sup>Other possible measures of *MP* include comparative levels of spending between the United States and

The use of this variable is consistent with the work of Dagobert Brito and Michael Intriligator on the international distribution of wealth: "...the distribution of wealth among nations and the distribution of the gains from trade is a function not only of markets and initial resources, as in the classical theory of international trade, but also of the power of the nations involved..." (1976, pp. 303-04). The expectation is that the existence and exercise of U.S. military power gives the United States and U.S. firms bargaining power and advantage in international economic transactions and that, consequently, it will have a positive relationship to the profit rate.

Table 1 shows the regression results. The coefficient for *MP* is positive and significant at the 5 percent level for a one-tailed test. In addition, all the variables have their expected signs and are at least weakly significant, except for  $\Delta RMP$ . Adding a Time trend slightly improves the results. These regressions suggest that increases in the use of U.S. military power have a direct and positive effect on the profit rate. In addition, there is evidence that military power and relative military spending have an indirect effect on the profit rate through the terms of trade. In several regressions of the terms of trade on various measurements of relative U.S. military power, positive and significant relationships were found (see my 1986 paper). The greater relative U.S. military power, the better the terms of trade. And with declining military power, we would expect a deterioration in the terms of trade.

In summary, the use of military power is positively related to the profit rate. The terms of trade are positively related to the profit rate. And relative U.S. military power and spending are positively related to the terms of trade. Military power has direct and indirect effects on the profit rate. The estimation

TABLE 1—MILITARY POWER AND THE PROFIT RATE, 1948–81, REGRESSION RESULTS<sup>a</sup>

Independent Variables	Profit Rate	
<i>CUMF</i>	.1112 (3.014)	.1539 (4.239)
$\Delta ULC$	-.0821 (-1.317) <sup>b</sup>	-.1096 (-1.779) <sup>c</sup>
<i>TOT</i>	.0383 (1.039) <sup>d</sup>	.0393 (1.457) <sup>b</sup>
$\Delta RMP$	-.0074 (-.349)	-.0084 (-.402)
<i>MP</i>	.0014 (1.707) <sup>c</sup>	.0014 (2.269)
Time		.0010 (3.092)
Constant	-.0459 (-.974)	-.1003 (-2.514)
Adjusted $R^2$	.3060	.5211
D-W	1.609	1.825

<sup>a</sup>T-statistics are shown in parentheses below each estimated coefficient. The Cochrane-Orcutt method was used to correct for autocorrelation.

<sup>b</sup>Significant, 10 percent one-tailed test.

<sup>c</sup>Significant, 5 percent one-tailed test.

<sup>d</sup>Significant, 15 percent one-tailed test.

of this model is consistent with the propositions set forth earlier—that international economic activity is important to capital accumulation and that military power has been a prop to the postwar social structure of accumulation. These results, consequently, are consistent with the interpretation that international economic considerations are important factors in promoting high levels of military spending in the United States in the postwar period. The pursuit of military power is partly motivated by international economic objectives. Military power and the goal of creating a stable international economic order dominated by the United States have been important parts of the postwar social structure of accumulation. Efforts to slow the arms race and military intervention must recognize these complex economic factors promoting military spending in the United States.

## REFERENCES

Adams, Gordon and Gold, David, *Defense Spending and the Economy: Does the Defense Dollar Make a Difference?*, Washing-

the Soviet Union, NATO, and other countries/regions; comparative force structures; or a composite of such factors. Other variables that might be useful in testing the hypothesis include economic and military assistance and their relation to the terms of trade and/or investment-trade patterns.

- ton: Defense Budget Project, 1987.
- Baran, Paul and Sweezy, Paul, *Monopoly Capital*, New York: Monthly Review, 1966.
- Becker, William H. and Wells, Samuel F., Jr., *Economics and World Power*, New York: Columbia University Press, 1984.
- Blechman, Barry M. and Kaplan, Stephen S., *Force Without War*, Washington: The Brookings Institution, 1978.
- Block, Fred L., *The Origins of International Economic Disorder*, Berkeley: University of California Press, 1977.
- Bowles, Samuel, Gordon, David M. and Weisskopf, Thomas W., "Power and Profits: The Social Structure of Accumulation and the Profitability of the Postwar U.S. Economy," *Review of Radical Political Economics*, Nos. 1&2, 1986, 18, 132-67.
- Brito, Dagobert L. and Intriligator, Michael D., "International Power and the Distribution of World Wealth," in Nake M. Kamrany, ed., *The New Economics of the Less Developed Countries*, Boulder: Westview Press, 1976.
- Cypher, James, "Capitalist Planning and Military Expenditures," *Review of Radical Political Economics*, No. 3, 1974, 6, 1-19.
- Dumas, Lloyd J., "Resource Diversion and the Failure of Conventional Macrotheory," *Journal of Economic Issues*, June 1983, 17, 555-64.
- Kindleberger, Charles P., *The World in Depression*, Berkeley: University of California Press, 1973.
- Meier, Gerald M., "Terms of Trade," in John M. Letiche, ed., *International Economic Policies and Their Theoretical Foundations*, New York: Academic Press, 1982.
- Melman, Seymour, *The Permanent War Economy*, New York: Simon & Schuster, 1974.
- Reagan, Ronald, "American Foreign Policy Challenges in the 1980s," *Weekly Compilation of Presidential Documents*, No. 2086, 1984, 1-6.
- Reich, Michael, "Military Spending and Production for Profit," in Richard Edwards et al., eds., *The Capitalist System*, 2nd ed., Englewood Cliffs: Prentice-Hall, 1978.
- Riddell, Tom, "Militarism: The Other Side of Supply," *Economic Forum*, Summer 1982, 13, 49-70.
- \_\_\_\_\_, "Military Spending and the Pursuit of Hegemony Under the First Reagan Administration," unpublished, 1985.
- \_\_\_\_\_, "Military Power, Military Spending and the Profit Rate, 1948-1981," unpublished, 1986.
- Schultz, George, "Restoring Prosperity to the World Economy," *Department of State Bulletin*, No. 2072, 1983, 64-68.
- Smith, R. P., "Military Expenditure and Capitalism," *Cambridge Journal of Economics*, March 1977, 1, 61-76.
- Congressional Budget Office, *Defense Spending and the Economy*, Washington: USGPO, 1983.

# THE ECONOMICS OF THE AGING OF THE BABY BOOM<sup>†</sup>

## The Baby Boom's Legacy: Relative Wages in the Twenty-First Century

By PHILLIP B. LEVINE AND OLIVIA S. MITCHELL\*

The economic impact of the large cohort born between 1946 and 1964 has been explored by several researchers. Analysis to date focuses mainly on the downward pressure on baby boomers' wages as their cohort entered the labor force (see R. B. Freeman, 1979; Louise Russell, 1982; Finis Welch, 1979). The present paper extends this literature by assessing the baby boom's impact on relative wages in the year 2020 when this generation will be the oldest segment of the workforce.

Several important public policy questions are addressed. First, will the changing demographic structure decrease the relative wages of prime-age workers? If so, there may be justification for social policy encouraging early retirement among those age 55+ to lessen downward pressure on prime-age workers' wages. A second question addressed is, how will the graying of the workforce affect teenage workers' wages? Because teens' wages and school attendance are linked, pay reductions may influence their investments in human capital and future earnings potential (R. G. Ehrenberg and A. J. Marcus, 1982). Finally, we investigate whether changing age structures are predicted to affect the female-male wage gap forty years hence.

The analysis uses national time-series data (1955-84) to estimate an econometric model of the demand for workers in eight different age-sex categories. Labor groups analyzed by sex are teens (ages 16-19), young workers (20-34), mature workers (35-54) and older workers (55+). Estimated coefficients are employed to predict changes in relative wages to the year 2020, when the youngest of the baby-boom group will be over age 55.

### I. Methodology and Data

Daniel Hamermesh and J. H. Grant (1979) recommend using a production function approach to compute how wages would change in response to changes in factor quantities. Rather than estimating a translog model directly, we estimate the coefficients in the relevant output share equations. In the empirical application below, cost shares will be utilized as the dependent variable since in competitive equilibrium they are equal to output shares.

Estimated coefficients are used to compute elasticities of complementarity and factor price elasticities. Elasticity variances are computed by applying the *delta* method.<sup>1</sup>

Coefficient estimates will also be employed in the policy simulation to determine the total effect of a changing labor force on relative wage rates. The effect of a quantity change ( $\% \Delta X_i$ ) on wages of labor subgroup  $i$  ( $\% \Delta W_i$ ) is computed as<sup>2</sup>

$$\% \Delta W_i \approx 1/S_i \left[ \sum_j \gamma_{ij} (\% \Delta X_j) \right] - \% \Delta X_i + \sum_j S_j (\% \Delta X_j),$$

<sup>†</sup>*Discussants:* David Bloom, Harvard University and NBER; James Poterba, Harvard University and NBER; John Hamor, U.S. Social Security Administration.

\*Department of Economics, Princeton University, Princeton, NJ 08544, and Department of Labor Economics, Cornell University, Ithaca, NY 14851, respectively. We are grateful to Ron Ehrenberg for comments on an earlier draft, Whitney Newey for statistical advice, Eileen Driscoll and Pam Rosenberg for computing advice, and Cornell University for research support. We remain solely responsible for views expressed herein.

<sup>1</sup>The derivation is available in our NBER working paper with the same title as this paper.

<sup>2</sup>The derivation of this formula is available in our NBER working paper.

where  $S_i$  is the share of the  $i$ th input to total cost and the  $\gamma_{ij}$  are estimated translog coefficients.

Like all production function models, the framework assumes that input supply changes are exogenously determined. We do not attempt to relax this assumption since there exist few instruments in time-series data. The likely effect of instrumenting has been shown to be negligible in a study by Ehrenberg and R. S. Smith (1987) though George Borjas (1986) finds that instrumenting alters a few of his findings.

The model is estimated with symmetry and homogeneity imposed.<sup>3</sup> Imposing these cross-equation constraints on the system of equations implies that disturbance terms may be correlated across equations. Thus the model is estimated using an iterative Seemingly Unrelated Regressions technique.

Estimation requires data on the quantity of each labor input, capital, and each input's share of total costs. All variables are annual national aggregates.<sup>4</sup> Derivation of employment, hours, weeks, and wage data is detailed in Freeman. Capital quantity and price data are taken from the MIT-Penn-SSRC (MPS) data bank.

## II. Elasticity Estimates

Table 1 presents statistically significant substitutes and complements within all labor categories.<sup>5</sup> Two conclusions emerge: 1) Most substitution occurs across gender for different age categories. Complementarity occurs across age groups for a given gender (with the exception of teenagers). 2) Older males are complementary with young males, and substitutable with female teens. Older females are substitutable with mature males.

TABLE 1—STATISTICALLY SIGNIFICANT  
COMPLEMENTS AND SUBSTITUTES<sup>a,b</sup>

Complements	Substitutes
<i>FT-MM</i> +4.24	<i>FT-MO</i> -7.99
<i>FY-FM</i> +3.86	<i>FT-FY</i> -7.80
<i>MY-MO</i> +1.07	<i>FT-FM</i> -7.22
	<i>FO-MM</i> -1.65
	<i>FY-MM</i> -0.81

<sup>a</sup>Elasticities are statistically significant at the 95 percent level.

<sup>b</sup>Elasticities of factor complementarity are ranked from highest to lowest. The variable definitions are *FT* = Female Teen (16-19); *MT* = Male Teen (16-19); *FY* = Female Young (20-34); *MY* = Male Young (20-34); *FM* = Female Mature (35-54); *MM* = Male Mature (35-54); *FO* = Female Older (55+); *MO* = Male Older (55+).

## III. Policy Simulation

To determine the impact on relative wages caused by the aging of the workforce, we apply the simulation formula above to our coefficient estimates and projections of how the entire age distribution is likely to change over time. Table 2 reports two projections of labor force patterns between 1985 and 2020 by age-sex subgroup obtained from data published by the Bureau of Economic Analysis (BEA) and the Social Security Administration (SSA). Both series are used in the empirical analysis below since the magnitudes differ due to different extrapolation methodologies. Both forecasts show the percentage of older workers will increase substantially as the baby boom ages. Predicted growth in female participation also implies larger changes for women than men. Changes in each labor group's wages are computed allowing capital to vary as predicted.<sup>6</sup>

Reported simulation results (see Table 3) indicate the predicted change in the wage of several labor subgroups between 1985 and 2020. If labor supply patterns behave according to projections, the evidence indicates

<sup>3</sup>A test for symmetry and homogeneity is not rejected at the 5 percent level. Tests for separability of labor from capital and consistent male and female aggregates are all rejected at conventional levels.

<sup>4</sup>A data appendix containing complete descriptive statistics is available in our working paper.

<sup>5</sup>Factor price and factor complementarity elasticities are available in our working paper.

<sup>6</sup>The value of capital stock for 2020 is imputed from a regression of actual capital stock from 1955 to 1984 on a trend variable.

TABLE 2—PROJECTED CHANGES IN LABOR SUPPLY  
BY AGE AND SEX: 1985 (actual)–2020

Demographic Group	Projections	
	BEA (1)	SSA (2)
Female Teen	21.5	19.5
Female Young	3.8	4.1
Female Mature	32.4	47.3
Female Older	54.5	92.0
Male Teen	16.5	18.4
Male Young	-9.9	-4.7
Male Mature	23.5	32.7
Male Older	46.1	81.2

Notes: Shown in percent. Col. 1 is the difference between the actual number of workers in that age/sex group in 1985 and BEA projections for 2020 (U.S. Department of Commerce, 1981). Col. 2 is the difference between actual number of workers in that age/sex group in 1985 and SSA projections for 2020 (U.S. Social Security Administration, 1983). The age groups are given in Table 1.

TABLE 3—THE BABY BOOM'S IMPACT ON WAGES:  
2020 vs. 1985

Average Predicted Wage Change for:	%Δ in Wage (BEA projection)	
	(1)	(2)
Older Workers	5.6	1.2
Mature Workers	7.8	4.4
Young Workers	6.1	5.7
Teen Workers	24.3	1.5
Female Workers	-10.3	-7.8
Male Workers	12.2	8.6

Notes: Shown in percent. Col. 1 is the BEA projection; col. 2 is the SSA projection. The age groups are given in Table 1.

that the aging of the workforce will have little effect on the wage distribution by age. While older workers' wages are predicted to increase 1.2 to 5.6 percent, prime-age workers' (mature and young) are predicted to increase a similar 4.4 to 7.8 percent. This finding is contrary to the notion that incentives for early retirement are needed to protect prime-age workers' wages.

However, when we consider males and females separately, we see that prime-age women will be hurt relative to older workers. The predicted increase of 1.2 to 5.6 percent

for older workers is contrasted with a 10.8 to 15.7 percent decrease for prime-age women. Female workers as a whole will also be hurt in comparison with male workers. While male wages are predicted to increase 8.6 to 12.2 percent, female wages are predicted to decrease 7.8 to 10.3 percent. This result is driven by prime-age workers: among this age group, men's wages are forecasted to increase 11 to 14.9 percent and women's to decrease 10.8 to 15.7 percent. As a result, the female-male wage gap will rise by the year 2020, *ceteris paribus*.

The analysis of teens remains inconclusive because of the large differences between SSA and BEA results.

#### IV. Conclusions

Coefficients from a translog production function are used to estimate demand elasticities and predict the relative wages of men and women in the year 2020. Our elasticity results indicate that, with the exception of teens, substitution occurs across gender and complementarity occurs across age groups for a given gender. Also, we find several interdependencies with older workers: older men are complementary with young men and substitutable with teenage women, while older women are substitutable with mature men.

The simulation results indicate that wages of prime-age workers will not deteriorate in relation to older workers as a result of the aging of the baby boom cohort. Conclusions for teens cannot be drawn. The general result does not hold for women, however. Prime-age women are predicted to lose in comparison with older workers and with men, increasing rather than reducing wage differentials by sex, *ceteris paribus*.

#### REFERENCES

- Borjas, G. J., "The Sensitivity of Labor Demand Functions to Choice of Dependent Variable," *Review of Economics and Statistics*, February 1986, 68, 58–66.
- Ehrenberg, R. G. and Marcus, A. J., "Minimum Wages and Teenagers Enrollment-Employment Outcomes: A Multinomial Logit

- Model," *Journal of Human Resources*, Winter 1982, 17, 39-58.
- \_\_\_\_\_ and Smith, R. S., "Comparable Worth in the Public Sector," in D. A. Wise, ed., *Public Sector Payrolls*, Chicago: University of Chicago Press, 1987.
- Freeman, R. B., "The Effect of Demographic Factors on Age-Earnings Profiles," *Journal of Human Resources*, Summer 1979; 14, 289-318.
- Hamermesh, D. S. and Grant, J. H., "Econometric Studies of Labor-Labor Substitution and Their Implications for Policy," *Journal of Human Resources*, Fall 1979; 14, 518-42.
- Russell, Louise B., *The Baby Boom Generation and the Economy*. Washington: Brookings Studies in Social Economics, 1982.
- Welch, F., "Effects of Cohort Size on Earnings: The Baby Boom Babies 'Financial Bust'," *Journal of Political Economy*, October 1979, 87, S65-97.
- U.S. Department of Commerce, Bureau of Economic Analysis, *BEA Regional Projections*, Washington: USGPO, July 1981.
- U.S. Department of Labor, Bureau of Labor Statistics. *Employment and Earnings*, Washington: USGPO, January 1986.
- U.S. Social Security Administration, *Actuarial Study No. 90: Economic Projections for OASDI Cost Estimates*, Washington: USGPO, 1983.



# The Baby Boom, Housing, and Financial Flows

By JOYCE MANCHESTER\*

The baby-boom generation is now in the prime of its young adult years, ranging from ages 23 to 41 with the biggest cluster around age 31. The behavior of this cohort as it swarms into the labor force, clamors for homeownership, and borrows to finance commodities as well as children has far-reaching effects on economic patterns in the United States. Twenty years from now, we anticipate continued influence on the economy as the baby boomers participate in the rituals of middle-age such as saving for retirement and paying off mortgages.

While recent research has examined baby-boom effects on housing demand, per capita income, and rates of return in a piecemeal fashion, none has done so within a well-specified general equilibrium model with emphasis on housing and financial markets through time. The importance of demographics in determining household formation and the demand for housing is discussed in George Sternlieb and James Hughes (1986) and Roger Craine (1983), but few behavioral adjustments are recognized. Crowding effects not unlike those of Richard Easterlin (1986) prove to be crucial to the model developed below, with changes in the housing market and in consumption-savings behavior ameliorating the hardship.

This paper examines cohort effects as well as macroeconomic effects arising from the baby boom in a simple three-period overlapping generations model in which equilibrium must be maintained in the markets for loanable funds and housing at all times. A more thorough presentation appears in my earlier paper (1986). Agents purchase housing financed by a long-term adjustable rate mortgage through the banking system, which also acts as the depository for savings. Equi-

librium requires that the supply of savings be sufficient to meet the demand for mortgages and that the supply of housing be sufficient to meet the demand for homeownership. In a closed economy under these conditions, the rate of interest and the price of housing are endogenous, and levels of consumption, savings, and welfare depend on those variables. A real shock, such as the baby boom, invokes behavioral responses in housing purchases and consumption. These in turn affect borrowing and lending, not only for the baby boomers but for other generations as well.

The model is examined under two very different assumptions regarding expectations: myopia and perfect foresight. Under myopia, agents anticipate that steady-state values will prevail in the future, regardless of current shocks. The baby boom causes a sharp decline in the real rate of interest and a sharp increase in the real price of housing for two periods before these values gradually return to preshock levels. Under perfect foresight, agents are aware that the baby boom will occur three periods prior to the event, and they understand the economy sufficiently to anticipate the paths of all relevant variables correctly. In this scenario, the real rate of interest rises just prior to and during the appearance of the baby boom, while housing prices increase when the baby boomers are young and middle-aged. Although it is difficult to compare social well-being under the two regimes, baby boomers fare least well in either case, while the generation to follow benefits most.

Emphasizing financial market flows and the housing market does omit some potentially interesting behavioral responses. The model is limited in that it neglects labor-market effects as well as lifestyle and fertility changes that might arise from shifting social and economic conditions. Indeed, labor-supply decisions are ignored, there is no

\*Department of Economics, Dartmouth College, Hanover, NH 03755.

capital other than housing, and individuals rather than families represent the agents of interest. These simplifications, while restrictive, do allow us to analyze the interaction between the financial market and the housing market more readily.

### I. The Model

Each generation consists of one agent who lives three periods, receives a share of the aggregate output related to his (or her) marginal product in each period, and must purchase a house financed with a long-term adjustable rate mortgage (*ARM*) at the start of the first period. No consumption loans are allowed. The owner sells the house and receives the proceeds at the start of the third period, but he continues to live in the house until his death. There are no bequests. Housing depreciates in each period, with new investment in housing a function of the price of housing,  $p$ , and the level of income,  $y$ . Both the current one-period rate of interest,  $r$ , and the expected rate determine the *ARM* payment to be made in the next period. Aggregate output is determined by a Cobb-Douglas production function that depends on the number of workers of different ages and their marginal products.

Agents maximize a Cobb-Douglas utility function in which real consumption and the quantity of housing appear explicitly. Consumption in the first period is a fixed percentage of first-period income. Each agent solves a dynamic programming utility-maximization problem to determine consumption in the second and third periods. Current expectations of  $r$ ,  $p$ , and  $y$  appear in the budget constraints. Each agent also maximizes utility with respect to housing. Housing demand depends positively on the relative weight placed on housing in the utility function, on future income, and on the expected selling price of housing. Interest rates and the current price paid for housing act as a negative influence through the size of mortgage payments.

The first equilibrium condition requires that housing demand be equal to housing supply. The supply of housing is composed

of the partially depreciated stock sold by old agents plus new investment in housing.

The second equilibrium condition combines two financial market equalities based on the mutual form of banking in which no change in net worth is allowed. First, flows into and out of the banking system must be equal. Inflows include the first payment on the mortgage loan negotiated last period, the second payment on the loan negotiated two periods ago, and savings deposits of the middle-aged. Outflows include the current housing loan and the savings withdrawal with accrued interest of the old generation. Second, assets must equal liabilities in each period. Assets of the bank include mortgage payments to be paid this period as well as the discounted value of the mortgage payment to be made next period. Liabilities are savings withdrawals and accrued interest. The assets-liabilities condition together with the inflows-outflows condition yield the financial market constraint.

The third equilibrium condition is the Cobb-Douglas production function. Conditions in the housing and financial markets, together with the production function, are solved simultaneously to find values of  $r$ ,  $p$ , and  $y$ . All other values are derived from these endogenous variables and the parameter values.

Special mention should be made of the way in which the rate of return on deposits,  $rd$ , is determined. Because current bank shareholders bear all risk of fluctuating asset values in this model and the only assets in the bank are *ARMs*, it is always true that the rate of return on deposits equals the previous rate of interest. Expectations of  $rd$  influence savings decisions, and realized values influence attained consumption of the elderly.

Parameters are chosen such that in the steady state, income is 1000 and one unit of housing is available to each generation. Utility function weights roughly correspond to a time preference parameter of 1.5 percent per year. Elderly individuals are in retirement and receive none of the current output, while young agents receive 25 percent and middle-aged agents 75 percent of the current

production. Housing depreciates 20 percent per period. In the steady-state solution,  $r$  equals 0.261 and  $p$  equals 165.84.

## II. Baby-Boom Effects in a General Equilibrium Model

The baby boom is modeled as a 20 percent increase in the number of agents born at  $t = 5$  only. When the baby-boom generation appears, it receives 25 percent of the aggregate income, which rises from 1000 to 1046 but must be shared by the larger number of agents. The middle-aged at  $t = 5$  benefit more from the rise in output since they receive 75 percent of this increased national product. Upon reaching middle-age, the baby boom cohort receives 75 percent of 1146, again divided among the larger number of agents.

### A. The Case of Myopic Expectations

First, let us examine the path of the economy when all agents believe that steady-state values will prevail in the periods ahead regardless of current economic conditions. For example, the calculation of the first *ARM* payment when the baby boom appears at  $t = 5$  is a function of current  $p$ ,  $h$ ,  $r$ , and the expected rate of interest next period, equal to the steady-state value under myopic expectations. In the following period, the realized value of  $r$  will not equal last period's expected value, implying that the second mortgage payment determined at  $t = 6$  differs from the first mortgage payment determined at  $t = 5$ . This affects savings by the middle-aged as well as the rate of return earned by the elderly. In addition, housing demand by the young is affected directly by the price they must pay to borrow funds.

At the time of the baby-boom shock, the rate of interest falls 26 percent while the price of housing rises 5 percent. The decline in the rate of interest seems counterintuitive at first glance, but it can be explained as follows. Increased demand for housing by the baby-boom cohort drives up the demand for mortgage loans, but this is more than offset by the increase in savings by the middle-aged who benefit from increased national output. The price elasticity of housing and

the increase in national income are not great enough to maintain per capita housing at one unit as the baby boomers purchase only 0.85 units per person.

As a result of these changes, the utility of elderly agents rises slightly and that of middle-aged agents increases even more. The old receive 5 percent more than anticipated when selling their house, while the mortgage payment and rate of return on deposits were determined prior to the shock and thus remain at their steady-state values. Those who are middle-aged benefit from the increase in aggregate income at  $t = 5$ . In addition, the second *ARM* payment to be made at  $t = 6$  is 5 percent lower than the steady-state value. The combination of higher income and lower mortgage payments allows the middle-aged to consume more as well as save more at  $t = 5$ . Their good fortune continues at  $t = 6$  as the price of housing received is 13 percent above the expected steady-state level. This capital gain together with increased savings more than offsets the lower level of  $rd$ , and third-period consumption is 5 percent above the steady-state level. The generation born just prior to the baby boom is better off than generations in the steady state.

Individuals in the baby-boom generation suffer more than any other agents due to crowding effects in the housing and financial markets and decreased income per capita. When young, they purchase only 0.85 units of housing at lower rates of interest, but the price per unit is 5 percent above the steady-state level. Mortgage payments are lower in the future, but lower utility is derived from housing. When middle-aged, they consume and save less per capita than the steady-state values due to reduced income only partially offset by decreased mortgage payments. Baby boomers suffer a small capital loss in selling their house and receive a much lower rate of return on deposits. Indeed,  $rd$  falls 62 percent between  $t = 6$  and  $t = 7$  as the stock of deposits on which the return is calculated is very large and the second mortgage payment made at  $t = 7$  is relatively small. As a consequence, per capita consumption for the baby-boom cohort when old is 16 percent below the steady-state level. Baby boomers do not fare well.

The generation born just after the baby boom is better off than any other. The rate of interest is extremely low at  $t = 6$  owing to the large supply of savings provided by the middle-aged baby boomers. The low cost of borrowing allows this generation to bid up the price of housing 13 percent above normal, invoking a supply response that is reinforced by increased output. Hence this generation enjoys 6 percent more housing than the steady-state level. The increased price and quantity of housing and the upward movement of the rate of interest imply mortgage payments that are slightly above the steady-state values. In middle age, less saving partially offsets the higher mortgage payment so that consumption is close to the steady-state level. In old age,  $rd$  soars to 13 percent above the steady-state level. This more than offsets the higher mortgage payment and capital loss incurred when the house is sold, and third-period consumption is slightly higher than in the steady state.

#### B. *The Case of Perfect Foresight*

In this scenario, agents become aware of the impending population bulge at  $t = 2$ , one full generation before the baby boom appears. Expectations adjust immediately so that expected values equal realized values. In the long run, neither the macroeconomic nor microeconomic effects turn out to be very different from those under myopia, but the immediate effects do differ.

While small movements in macroeconomic variables appear as soon as the upcoming population bulge is announced, the most interesting economic effects occur close to the appearance of the baby boom. In contrast to the decline in  $r$  at  $t = 5$  under myopic expectations,  $r$  rises 17 percent at  $t = 4$  and rises an additional 16 percent at  $t = 5$  under perfect foresight. From its peak at  $t = 5$ , the rate of interest falls 70 percent at  $t = 6$  and then overshoots the steady-state level before gradually declining toward it. The behavior of the price of housing is similar to that under myopic expectations, although the magnitude of the movements is smaller. Welfare of each generation is very similar in the two scenarios, despite the dif-

ferences in the rate of interest and in the resulting patterns of consumption and saving.

The behavior of the rate of interest is explained readily by considering the supply and demand of loanable funds. At  $t = 4$ , the expectation of increased  $rd$  and capital gains at  $t = 5$  causes the supply of savings to shift back. At the same time, the demand for loanable funds shifts back by a small amount as the price of housing falls slightly. Together these shifts imply an increase in  $r$  and a small decline in loanable funds. The picture becomes slightly more complex at  $t = 5$ . Both the supply and demand for loanable funds shift out. Higher income owing to the population bulge yields more savings by those who are middle-aged. The demand for loanable funds, consisting of the new loan and the discounted value of the mortgage payment to be received next period, shifts out by a lesser amount. Elasticities are such that both the rate of interest and the flow of savings increase at the time of the baby boom shock.

When the baby boomers reach middle age, the aggregate supply of savings rises to 18 percent above its steady-state value. In part this reflects the unit income elasticity of savings of the Cobb-Douglas utility function. The new mortgage loan and the present value of the second mortgage payment rise as well, but the supply effect dominates and  $r$  plummets 70 percent at  $t = 6$ . Both the price and quantity of housing increase substantially at  $t = 6$  as the young agents bid for housing given the favorable financing conditions.

The welfare of baby boomers is again the lowest of any generation. Per capita consumption is below steady-state levels 13 percent when young, 4 percent when middle-aged, and 16 percent when old. Per capita housing again lies 15 percent below the steady-state level. Behavioral changes only partially ameliorate the crowding effects of the 20 percent population bulge.

Just as in the case of myopic expectations, those born just after the baby boom enjoy the highest utility. The cost of borrowing at  $t = 6$  is very low, while the price and quantity of housing are both quite high. Mortgage

payments are only slightly above the steady-state level. First-period consumption is 15 percent above the steady-state level as the middle-aged baby boomers produce more output, some of which goes to the young. Second-period consumption is equal to the steady-state value, while savings is slightly lower in anticipation of slightly higher  $rd$  to be received at  $t=8$ . Third-period consumption lies just above the steady-state level as the decline in  $p$  is offset by the larger quantity of housing sold.

From  $t=7$  forward, economic conditions are not very different from the steady state. The slowly depreciating stock of housing remains above one unit for some time, while its price slowly moves up to the preshock level. The rate of interest is within 1 percent of the preshock level by  $t=10$ , and the flow of loanable funds is very close to its steady-state level as well. Those generations born at  $t=7$  or later are slightly better off than those in the steady state owing to the slightly larger quantity of housing.

### III. Sensitivity of Results to Choice of Parameter Values

Quantitative as well as qualitative movements in the endogenous variables may be sensitive to the choice of parameter values in the simulations discussed here. Experiments using alternative parameter values for the rate of time preference, depreciation of the housing stock and specification of the housing investment function, the share of income going to young vs. middle-aged workers, and the elasticity of substitution between young and middle-aged workers in the production function have been performed. Only in the case of a very small elasticity of substitution do the qualitative results differ from those of the base case discussed in Section II.

The Cobb-Douglas production function implies an elasticity of substitution equal to one. A CES production function with this elasticity equal to three leads to wider swings in  $r$  while preserving the same qualitative trends. Specifying the elasticity of substitution to be extremely small, however, yields rather different results. In contrast to de-

clines in  $r$  under myopic expectations in the base case, the CES production function with an elasticity of substitution of one-third causes the rate of interest to decline 70 percent at  $t=5$  and rise 269 percent at  $t=6$ . Saving is lower at  $t=6$  as the income received by the middle-aged baby boomers falls more sharply. In the case of perfect foresight, movements in the rate of interest are reversed both at  $t=5$  and  $t=6$ . In the base case,  $r$  rises 16 percent at  $t=5$  and falls 70 percent at  $t=6$ . The smaller elasticity of substitution implies that  $r$  falls 59 percent and rises 77 percent at  $t=5$  and  $t=6$ . Wages adjust such that baby boomers receive less income when young and middle-aged than in the base case, while the generations on either side receive more. Baby boomers also receive a lower rate of return on their savings when old. These two factors imply that saving falls precipitously at  $t=6$  following a substantial increase at  $t=5$ . Shifts in the demand for loanable funds reinforce these trends.

### IV. Concluding Remarks

This paper emphasizes the importance of financial market effects in a general equilibrium model with housing when a baby boom occurs. Agents in this three-period overlapping generations model borrow to finance the purchase of a house when young and make mortgage payments over the remainder of their lifetimes. They save when middle-aged to provide for consumption when old, and the rate of return earned on deposits together with the price at which housing is sold influence third-period consumption and hence utility.

Welfare of individuals and the path of the economy were examined under the assumptions of myopic expectations and perfect foresight. Utility of corresponding generations is very similar under either scenario. Baby boomers suffer more than others due to crowding effects in the housing and financial markets and decreased income per capita. The generation born just after the baby boom is better off than any other as its members enjoy more housing per capita

combined with low borrowing costs. Whereas the rate of interest falls at the time of the baby boom under myopic expectations, however, it rises just prior to and during the appearance of the baby boom under perfect foresight. This difference is attributed to the relative size of shifts in the supply and demand for loanable funds together with adjustments in housing demand, mortgage contracts, and planned consumption caused by the baby boom. These results were shown to be robust to several changes in parameters, with different paths for the rate of interest occurring only in the case of a very small elasticity of substitution between young and middle-aged workers in the production function. Adjustments in the loanable funds market and in consumption-savings behavior may well ameliorate some of the demand pressures on the housing market emphasized

in other studies of baby-boom effects on the economy.

## REFERENCES

- Craine, Roger, "The Baby Boom, the Housing Market, and the Stock Market," *Federal Reserve Bank of San Francisco Economic Review*, No. 2, Spring 1983, 6-11.
- Easterlin, Richard A., "Easterlin Hypothesis," unpublished manuscript, University of Southern California, January 1986.
- Manchester, Joyce, "The Baby Boom, Housing, and Loanable Funds," Working Paper 86-6, Dartmouth College, December 1986.
- Sternlieb, George and James W. Hughes, "Demographics and Housing in America," *Population Bulletin*, January 1986, 41, 3-34.

# Social Security Benefits and the Baby-Boom Generation

By TABITHA A. DOESCHER AND JOHN A. TURNER\*

Currently there are 3.7 workers supporting each recipient of benefits from the Social Security Old-Age and Survivors Insurance (OASI) program. By 2030, when the last of the baby-boom generation reaches retirement age, this ratio is expected to fall to 2.2 workers per beneficiary.<sup>1</sup> Because the U.S. Social Security system has been a pay-as-you-go system, the expected decline in this ratio has fueled considerable debate over the future of Social Security. In particular, many members of the baby-boom generation are concerned about the level of Social Security benefits they will receive when they retire.

The Social Security Administration (SSA) has examined this issue (1986). Using an actuarial approach, the SSA concludes that the system, as it is currently structured, is in actuarial balance and that baby boomers can expect to receive retirement benefits that are generous by today's standards. However, a major weakness of the SSA approach is that it ignores political forces and the potential impact of these forces on benefit levels.

This paper addresses this weakness by drawing upon voting and special-interest group models to develop an economic, as opposed to an actuarial, approach for predicting OASI benefits. Following a discussion of the SSA's projections of future benefit levels, the paper presents a theoretical model which postulates that OASI benefits are determined in the political market. In this model, changes in the old-age dependency ratio (the ratio of OASI beneficiaries to covered workers) affect the level of individual benefits in two opposing ways.

First, as the number of retirees increases relative to the number of workers, the retirees acquire additional political power which can be used to raise individual benefits. However, at the same time, the cost of increasing individual benefits is greater, and this may lead to a lowering of individual benefits. The net impact is indeterminate and must be examined empirically.

Based on empirical estimates of the relationship between benefit levels and age dependency ratios, and based on demographic and economic projections used by the Social Security Administration, the paper predicts the level of Social Security benefits the baby-boom generation can expect to receive. These predictions are compared to the OASI actuarial projections of future benefits.

## I. The SSA Projections

The Social Security Administration projects OASI benefits and tax payments 75 years into the future. These projections are actuarial projections and are based on legislated payroll tax rates, legislated OASI benefit formulas, and projected demographic and economic conditions. The SSA actuaries develop four projections based on varying degrees of optimism about future demographic and economic conditions. The following discussion is based on the less optimistic of the intermediate projections (alternative II-B).

### A. Benefits

The OASI benefits are one of the primary outputs of the SSA actuarial projections. The most recent SSA projections indicate that the OASI benefits received by the baby-boom generation will be considerably higher in real terms than those received by workers who retired in 1986. For example, according to these projections, a male worker

\*Office of Business and Economic Research, Oklahoma State University, Stillwater, OK 74078, and Pension and Welfare Benefit Programs, U.S. Department of Labor, Washington, D.C. 20210, respectively. The opinions expressed in this paper do not necessarily represent those of the U.S. Department of Labor.

<sup>1</sup>Estimates of this ratio range from 2.6 to 1.8. See SSA (1986, Table 30).

born in 1955 and retiring in the year 2020 at age 65 will receive before-tax benefits that are 60 percent higher in real terms than the benefits received by a 65-year-old male who retired in 1986. Similarly, a female worker born in 1955 can expect benefits 72 percent higher than those being received by a recently retired female beneficiary.

The after-tax comparisons are only slightly less favorable. In 1986, 2.0 percent of OASI benefits were paid back to the government in federal income tax payments. The OASI actuaries project that this figure will rise to 4.47 percent by 2020. The effect of taxation is to reduce the real increase in retirement benefits from 60 percent to 58 percent for males and from 72 percent to 70 percent for females.

An additional consideration is the effect of changes in the retirement age on the level of benefits. In 1983, the Social Security Act was amended to gradually increase the age at normal retirement, beginning in the year 2000. This legislated postponement in the normal retirement age is effectively a reduction in benefits for workers retiring before age 67. When this change is taken into account, the outlook for baby-boom generation retirees becomes less favorable, though real benefits are still expected to be considerably higher than in 1986. After adjusting for the normal age of retirement, real OASI benefits at age 65 are projected to be 43 percent higher for males retiring in 2020 than for males retiring in 1986. Real after-tax benefits are expected to be 42 percent higher. For females, real benefits are expected to be 53 percent higher, while real after-tax benefits are projected to be 52 percent higher.

Real benefits for family units (as opposed to individuals) are more difficult to compare. The preceding comparisons are accurate for family units only if the family units are composed of a single Social Security recipient. Any increase in the proportion of the female population receiving Social Security benefits in their own right rather than as dependents of a male will cause the preceding comparisons to understate the improvement in the economic status of Social Security recipients.

## B. Replacement Rates

Given a life cycle framework, comparisons with preretirement consumption levels may be more relevant for determining well-being in retirement than are intergenerational comparisons. Therefore, the SSA calculates replacement rates, that is, ratios of annual OASI benefits at retirement to annual earnings in the year preceding retirement. The primary conclusion from the projections of the Social Security Administration is that the replacement rates for baby-boom generation retirees will be basically at the same level as replacement rates for workers retiring in 1986 (SSA, 1986 *Annual Report*).

The replacement rates computed by the SSA, however, do not reflect the benefit reduction that will occur when the age at normal retirement is increased. The ratio of retirees' benefits (adjusted for retirement at age 65) to the average earnings of covered workers indicates that male Social Security recipients in 2020 (born in 1955) will be 13 percent worse off than were male Social Security recipients in 1986. Similarly, female Social Security recipients in 2020 will be 6 percent worse off than their more elderly counterparts.

## C. Discussion

In summary, the projections of the SSA indicate that members of the baby-boom generation who retire in 2020 at age 65 will receive OASI retirement benefits approximately 40 to 50 percent higher in real terms than those of their parents' generation, but will have replacement rates 6 to 13 percent lower. The SSA calculates these expected benefit levels based on economic and demographic projections (for example, age distributions, life expectancy, and future earnings). The SSA actuaries apply the current provisions of the Social Security Act to these projections to estimate future OASI disbursements. While this approach provides many valuable insights for policymakers, it fails to consider how political forces may affect the Social Security Act and therefore affect the level of individual benefits.



There are a number of reasons to expect Congress to amend the Social Security Act over the next 30 or 35 years. First, there is considerable historical precedent: Congress has enacted major changes in the Social Security Act seven times in the past 30 years.<sup>2</sup> Second, the current provisions of the Social Security Act will result in the buildup of a huge surplus in the OASI trust fund during the 20 years preceding the retirement of the baby-boom generation. While this enables the program to fund baby-boom benefits at projected levels, maintaining the integrity of this surplus will be a challenge. For example, as the surplus is built up, individuals who are working may exert political pressure to reduce the payroll tax rate. At the same time, individuals who are retired (and who are members of pre-baby-boom generations) may argue that the surplus should be used to increase current benefits. If either of these events occur, the surplus will be reduced and, unless further changes are enacted, baby-boom retirees will be worse off than SSA projections indicate.

For these reasons, it is important to understand the forces that determine individual benefit levels. The remainder of this paper presents a political model of the determination of individual benefit levels. Based on this model, the paper predicts the level of OASI benefits the baby-boom generation can expect to receive. These predictions are then compared to the OASI actuarial projections of future benefits.

## II. The Model

The determination of the level of Social Security benefits can be viewed as the outcome of forces in the political market. Using this approach, the political process is analyzed as taking place in a market where, similarly to the money-exchange market, forces of demand and supply determine the equilibrium shadow prices and quantities of commodities. In this case, the commodities in the political market are transfers to ben-

eficiaries, expressed in quantity units of one dollar of transfers per beneficiary. The shadow prices are the per capita costs to taxpayers of the government making one additional dollar of transfers per beneficiary.

There are two government transfer programs in this political market. One program provides individual Social Security benefits of  $B$  to  $N_1$  retirees, while the other furnishes individual education benefits of  $Z$  to  $N_3$  youth beneficiaries. The remainder of the section will focus on the former program.

The per capita marginal cost (shadow price)  $p_B$  of Social Security benefits  $B$  per retiree is the increase in the individual worker's tax payments  $T$  with a marginal increase in benefits per beneficiary  $B$ :

$$(1) \quad p_B = dT/dB.$$

Assuming for the moment only one government program (Social Security), financed by a flat rate tax  $t$  paid by  $N_2$  taxpayers on income  $Y_2$  to fund Social Security benefits  $B$  for  $N_1$  beneficiaries, the pay-as-you-go budget constraint is

$$(2) \quad N_2 t Y_2 = N_1 B.$$

Assuming no behavioral reactions to changes in benefits or taxes, the shadow price of Social Security benefits is simply the old-age dependency ratio:

$$(3) \quad p_B = N_1/N_2.$$

The "suppliers" of transfers are individual taxpayers. The supply function is a marginal cost function which indicates the marginal per capita cost to taxpayers of the government providing additional transfers per beneficiary. The marginal cost function depends on the rate at which tax payments per taxpayer are translated into benefits per beneficiary.

The "demanders" in the model are voters, and the aggregate demand function depends on the views of different voter groups (beneficiaries and nonbeneficiaries) as to the

<sup>2</sup>The Act was amended in 1956, 1960, 1965, 1967, 1972, 1977, and 1983.

appropriate levels of transfers. The demand function is affected by politics, lobbying, majority voting, and the altruism or nonaltruism of different groups. This model thus maintains the usual taxonomy separating forces affecting marginal cost and forces affecting utility.

The two primary models of public choice, voting and special-interest group models, reach opposite conclusions concerning the effect of the large baby-boom generation on the level of its Social Security benefits. Voting models (see James Buchanan and Gordon Tullock, 1962) are generally interpreted to predict that larger groups obtain greater benefits per member because voting power increases with group size. Special-interest group models (see George Stigler, 1971) predict the opposite: small groups obtain higher benefits per member because the cost to the taxpayer of providing benefits increases with group size. In addition, small groups may be more efficient in obtaining financial support from their members for lobbying.

The political market model incorporates both the voting and special-interest group models. Special-interest group models are incorporated in the supply side: the larger is a group, the greater is the marginal cost of providing benefits to the group. Voting models are incorporated in the demand side: the larger is a group, the greater is its influence on (demand for) the level of benefits provided.

### III. Predicting OASI Benefits

The theoretical model developed in Section II can be used to predict OASI benefit levels given economic and demographic projections. The equilibrium equation for Social Security benefits is

$$(4) \quad \bar{B} = B(N, P, Y_2, X),$$

where  $N$  = the old-age dependency ratio (the shadow price of Social Security benefits paid to retirees),  $P$  = the youth dependency ratio (the shadow price of educational benefits paid to youths), and  $X$  is the expenditure on Social Security lobbying per nonbeneficiary lobbied;  $Y_2$  was defined earlier. The total dif-

ferential, expressed in elasticity form, is

$$(5) \quad E(\bar{B}) = \epsilon_{BN}E(N) + \epsilon_{BP}E(P) + \epsilon_{BY_2}E(Y_2) + \epsilon_{BX}E(X),$$

where  $E$  is the logarithmic differential operator. By using the SSA's economic and demographic projections and by working with estimates of the price and earnings elasticities, equation (5) is used to predict average OASI benefits for the 1955 birth cohort assuming retirement at age 65 (in 2020).

According to the SSA's intermediate projection (alternative II-B), between 1986 and 2020 the old-age dependency ratio is expected to increase by 52.2 percent, the youth dependency ratio is expected to decline by 13.2 percent, and the real income of covered workers is expected to increase by 63.5 percent. The increase in the old-age dependency ratio reduces benefits, while the decrease in the youth dependency ratio and the increase in real income raise benefits.

Empirical estimates of both the price elasticities can be derived from Turner (1984). The elasticity of Social Security benefits with respect to the old-age dependency ratio ( $\epsilon_{BN}$ ) is estimated to be  $-2.0$ , while the elasticity of benefits with respect to the youth dependency ratio ( $\epsilon_{BP}$ ) is estimated to be  $-1.7$ .<sup>3</sup> The earnings elasticity is assumed to equal 1.0.

Substituting these values into equation (5) yields a prediction that real Social Security benefits for the 1955 birth cohort (retiring at age 65 in 2020) will be 18.5 percent lower than they are for the cohort retired in 1986. The impact of the 52 percent increase in the old-age dependency ratio outweighs the effect of the 63 percent increase in income and the smaller decrease in the youth dependency ratio. In contrast, the SSA actuaries project that real benefits will be 42.1 percent higher.

If the effect of the old-age dependency ratio on individual Social Security benefits is nonlinear, with increasingly higher depen-

<sup>3</sup>These are large elasticities, with the old-age dependency elasticity implying a decrease in total Social Security benefits with an increase in population.

dependency ratios having decreasing effects, the point elasticity overestimates the true effect which would be calculated with an arc elasticity. Because of this possible source of overestimation and because of the size of the discrepancy between the SSA projections and the projections developed in this paper, the remainder of this section examines the sensitivity of the predictions to various price elasticity levels. Throughout this analysis, both price elasticities are assumed to be negative (consistent with the empirical evidence). The value of  $\epsilon_{BN}$  is assumed to range from  $-0.4$  to  $-2.0$  (its estimated value), while the value of  $\epsilon_{BP}$  is assumed to range from  $0$  to  $-1.7$  (its estimated value). The earnings elasticity is assumed to equal  $1.0$  and is not varied. Since the weighted average of income elasticities must equal one, that value is a natural choice when there is no evidence to the contrary.

The first set of calculations identifies the elasticities that are consistent with the projections of the SSA actuaries. Making the plausible assumption that the effect of old-age dependency on retired worker benefits is larger than the effect of youth dependency, the OASI actuarial projections imply that a 1 percent increase in old-age dependency would, *ceteris paribus*, reduce real retired worker benefits by roughly 0.5 percent. Thus, the official SSA projections imply a negative effect of the baby-boom generation on the level of individual Social Security benefits. However, this elasticity is considerably smaller than the  $-2.0$  value estimated by Turner (1984).

If a larger negative effect of old-age dependency is assumed, the predicted level of OASI retired worker benefits is reduced. For example, if it is assumed that a 1 percent increase in old-age dependency reduces old-age benefits by 0.7 percent rather than 0.5 percent, the growth of real retired worker benefits between 1986 and 2020 is reduced

from 42 percent to 30 percent (*ceteris paribus*). This implies that the OASI projections exceed the predicted level of benefits by almost 10 percent. If the SSA projections are overstated by 10 percent, then OASI retirement benefits will be approximately 28 percent higher in real terms for the baby-boom generation than for their parents and replacement rates will be roughly 22 percent lower.

In closing, it should be emphasized that the projections developed in this paper are based on a pay-as-you-go financing of Social Security OASI. The theoretical model, which forms the basis for the predicting equation, does not explicitly allow for the buildup of a large trust fund. In addition, the empirical estimates from Turner (1984) are for a period during which there was little change in the trust fund, presumably because the changes in the old-age dependency ratio were not sufficiently large to affect this fund. Thus the estimates presented in this paper suggest that if historical patterns continue, the benefits received by the baby-boom generation will be significantly lower than those projected by the Social Security Administration.

## REFERENCES

- Buchanan, James M. and Tullock, Gordon, *The Calculus of Consent*, Ann Arbor: University of Michigan Press, 1962.
- Stigler, George J., "The Theory of Economic Regulation," *Bell Journal of Economics*, Spring 1971, 2, 3-21.
- Turner, John A., "Population Age Structure and the Size of Social Security," *Southern Economic Journal*, April 1984, 50, 1131-46.
- Social Security Administration, *1986 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Disability Insurance Trust Funds*, Washington: USGPO, 1986.

## THE FEMINIZATION OF POVERTY<sup>†</sup>

### Child Support Payments: Evidence from Repeated Cross Sections

By ANDREA H. BELLER AND JOHN W. GRAHAM\*

Single-parent female-headed families are a large and growing proportion of all families and comprise a disproportionately large share of the poverty population. However, among these father-absent families, those who receive child support payments have higher incomes and lower poverty rates. Unfortunately, many eligible women have no child support awards and many with awards do not receive full payment. With low earnings in the labor market and difficulty finding inexpensive quality child care, many mothers have no choice but to rely on the welfare system.

The incidence of poverty is far greater among single-parent families headed by women than among other types of families. Of the 8.8 million women with children present and father absent in 1985, 2.8 million had incomes below the poverty level.<sup>1</sup> The absence of child support is a significant contributory factor. The poverty rate in 1985 for women not having child support awards was 49 percent; for those who did, it was 21 percent. Among those women awarded child

support, the poverty rate was 18 percent if they received payments and 27 percent if no payments were received. As these figures suggest, securing more child support from absent fathers should help to alleviate poverty among single-parent families and may help to reduce the welfare rolls.

Recent interest in improving the nation's child support enforcement system was sparked by the 1980 Census Bureau's release of alarming statistics on nonpayment rates. These statistics showed that, in 1979, only 59 percent of mothers with children from absent fathers had child support awards. Among those with an award and due payment in 1978, only about half received full payment—approximately one-quarter received partial payment, while one-quarter received no payment at all. Moreover, payments that were made were often irregular. The average amount of child support (conditional upon receiving support) was only \$1800 for an average of nearly two children.

The situation was considerably worse than average for black mothers and never-married mothers. Only about one-third of black mothers had child support awards and among those that did, the award amount was about 18 percent lower than for nonblacks (see our 1986 article). Only 11 percent of never-married mothers had awards. Award amounts were less than half as much as for divorced mothers.

#### I. Trends in Child Support Payments

While a minority of women are receiving child support, an upward trend may be underway. Examining trends in award rates, receipt rates (the proportion of women due child support who received something), and amounts received, shows a mixed picture.

<sup>†</sup>*Discussants:* Irwin Garfinkel, University of Wisconsin-Madison; Sheldon Danziger, University of Wisconsin-Madison.

\*University of Illinois, 274 Bevier Hall, 905 S. Goodwin Ave., Urbana, IL 61801 and Rutgers University, 360 Martin Luther King, Jr. Blvd., Newark, NJ 07102 and NBER, respectively. This research was supported by NIH grant no. HD19350-04. The opinions are strictly our own. We are grateful for the excellent research assistance of Seung Sin Chung and Sanghee Cha.

<sup>1</sup>Except where noted, the statistics are taken from Census Bureau reports on data from the March/April Match Files of the 1979, 1982, 1984, and 1986 *Current Population Surveys* (CPS) (U.S. Census, 1981, 1985, 1986, and 1987, respectively).

The child support award rate fell slightly from 59 to 58 percent between 1979 and 1984. The entire decline occurred between 1982 and 1984. The rate rose to 61 percent in 1986. The award rate did not decline among blacks, and rose among the never-married. Interestingly, groups with the lowest award rates appear to have made the most progress. Among blacks, the award rate increased from 29 percent in 1979 to 34 percent in 1982, where it remained in 1984—it rose to 36 percent in 1986. The award rate also increased continuously among the never-married from 11 percent in 1979 to 18 percent in 1984 and 1986. Thus, there appears to have been some change in the composition of the population awarded support.

In sharp contrast to the decrease in the award rate, the reciprocity rate increased between 1978 and 1983, from 71.6 to 76.0 percent. Again the entire change occurred between 1981 and 1983. (The rate decreased slightly in 1985.) Again, there are racial differences. The steady increase in award rates among blacks was accompanied by a steady increase in reciprocity rates, from 63 percent in 1978 to 69 (72) percent in 1983 (1985). By contrast, among the never-married, while award rates increased, reciprocity rates first fell substantially, from 81 to 63 percent, then rose to 76 percent in 1983 and 1985.

The mean amount of child support received (conditional upon receiving support) in constant dollars declined by 16 percent between 1978 and 1981, but then remained constant to 1983. The decline in real dollars received was accompanied by an even larger 19 percent decline in real dollars due. That decline was not terribly surprising because it occurred during a period of extremely high inflation (the Consumer Price Index rose about 40 percent). Rising prices erode the real value of child support awards made in earlier years, since awards typically do not have automatic adjustment clauses. What is surprising is that the real value of new awards decreased even more (see our 1986 paper). This suggests that the decline in award amounts, and related declines in receipts, might be attributable to compositional changes in the population of women awarded support and not solely to inflation. As additional evidence to this effect, the real

value of child support received fell 12 percent between 1983 and 1985, a period of relatively little inflation. Since real dollars due fell by less, part of the explanation for the declines in child support receipts must lie elsewhere.

## II. Analysis of Trends

The purpose of this paper is to investigate the reasons for these changes in child support between 1979 and 1983 by analyzing the micro data on which the *CPS* reports cited above are based.<sup>2</sup> Some change over this period could be expected due to changes in the underlying determinants of child support such as compositional changes in the relevant population, changes in economic conditions, in child support laws, and in public awareness concerning the problem of nonsupport.

Federal initiatives to establish a national child support enforcement program began in 1975 with the passage of Title IV-D of the Social Security Act. This program assists states in establishing paternity, locating absent parents, establishing child support obligations, and enforcing such obligations. Since 1975, many states added new child support laws to their existing ones. Further, total IV-D expenditures grew steadily from \$400 million in 1979 to \$700 million in 1983 (U.S. Office of Child Support Enforcement, 1983, p. 42). Finally, increased public awareness about the problem of nonsupport followed the release in 1980 of results from the first national census survey on child support.

To assess the relative importance of these various factors that have contributed to changes in child support over time, it may be useful first to review the determinants of child support at a single point in time. Our 1985 paper hypothesized that child support awards depend upon the anticipated needs of the custodial parent, the long-run ability of the absent parent to pay support, and state laws governing marital dissolution, paternity establishment, and child custody and support. Given an award, its receipt

<sup>2</sup>We examine trends through 1984 only because the 1986 *CPS* micro-data file has not yet been released.

depends upon the custodial parent's immediate needs, the obligee's current ability and desire to pay, and the state's enforcement program. While neither needs nor ability to pay can be observed directly in *CPS* data, we have argued that both can be proxied by socioeconomic characteristics of the mother (including race, ethnicity, education, age, number and ages of children, and current marital status). State differences in child support laws are proxied by a set of region dummy variables.

We focus our empirical analysis upon two measures of receipts, given those changes observed in the composition of the population awarded support and in the award amounts between 1979 and 1984. We use maximum likelihood probit to estimate the probability that a woman receives support in a given year (in aggregate, the receipt rate) and ordinary least squares to estimate the dollar value of her receipts, given she receives something. We estimate both the structural equations that include the value of the award as an independent variable, and reduced-form equations that replace the award amount with its underlying determinants. Each equation is estimated first with 1978 *CPS* data and then with 1983 data. (Complete regression results and other tables appear in an appendix available from the authors upon request.)<sup>3</sup>

### III. Regression Results

Many of the same factors significantly affect the probability of receiving child support in 1983 as in 1978. Older, more educated, and nonblack women are more likely to receive support, perhaps because their children's fathers (who likely share similar characteristics) are more able to pay. Fathers are also more likely to pay support if their award settlement was voluntarily

agreed to, if their children are older, or the more recent their marital disruption. In 1978 but not 1983, never-married mothers due child support (a small and probably select group) are more likely to receive it than ever-married mothers. In the structural receipt rate equations, each additional \$1000 of support due in 1978 (1983) increases the probability of receiving something by 3 (2) percent. In the reduced-form equations, the coefficients of the other variables are larger than in the structural (since they measure both receipt and award responses), but their statistical significance is little changed.

As expected, many of these same variables (race, age, education, voluntary agreements, and years since divorce/separation) are found to affect the amount of support received (conditional upon receiving support) in 1978 and 1983 (in constant 1983 dollars). In addition, family size is significant: each additional child raises 1978 receipts \$242 in the structural equation, and \$784 in the reduced form when award amounts are not controlled for directly. In the structural equation, each \$1000 increase in support due is found to raise receipts \$686 in 1978 and \$912 in 1983.

While we have seen above that child support receipts changed between 1978 and 1983, so did many of their underlying determinants. For example, this period saw large increases in the fraction of black and never-married women awarded child support, and decreases in average number of children and in real dollars of child support due. What change in receipts could thus be expected based solely on the changes in the socioeconomic characteristics of the population due child support? In addition, how much of the change in receipts can be attributed to the increased public awareness of child support and the new laws and enforcement techniques added by states during this period?

### IV. Decomposition Analysis

To answer the above questions, we employ a technique widely used in the discrimination literature known as "means-coefficients analysis." This allows us to decompose the 1978 to 1983 change in the receipt rate (or in

<sup>3</sup>Variables included in all regressions are mother's age, education, number of children, years since marital disruption, and dummy variables for her college graduation, children aged 6 to 17, marital status (never-married, remarried, separated, separated longer than two years), region (Northeast, South, Northcentral), SMSA, central city, black, Hispanic, and voluntary child support awards.

dollars received) into its explained and unexplained components. It can be shown that the explained portion equals the (1978 to 1983) change in the means of the independent variables evaluated at (multiplied by) the 1978 coefficients, while the unexplained portion equals the change in the coefficients evaluated at the 1983 means. In other words, the former represents the change that could be expected on the basis of observed changes in socioeconomic characteristics given laws and attitudes prevailing in 1978, while the latter represents the changes attributable to changes in child support laws and social attitudes themselves between 1978 and 1983.

In our regression sample of women due child support, the receipt rate (adjusted by a CPS weight to reflect population estimates) rose 4.52 percentage points between 1978 and 1983. Based upon the reduced-form estimates (that exclude child support due), we can explain a rise of 1.95 percentage points over this period with observed changes in the socioeconomic characteristics of the population due support. This leaves 57 percent of the actual increase unexplained. Based upon the structural equation, we would have expected a 1.04 percentage point fall in the receipt rate, leaving the entire increase unexplained. This occurs because the positive impact of some changes is more than offset by the negative impact of the 18 percent decline in real dollars due over this period of high inflation.

Some of the factors positively associated with the likelihood of receipt, in both structural and reduced forms, increased between 1978 and 1983. In order of their contribution to the increase in expected receipts, they are educational attainment (especially the fraction of college graduates), the proportion of women with voluntary (as opposed to court-ordered) awards, the proportion never-married, the proportion with older children, and the proportion living in the South. In addition, there was a decrease in the proportion of women separated longer than two years, a factor negatively associated with the likelihood of receipt. Partly offsetting these changes were two which tended to reduce the receipt rate. These were increases in the average length of time since the marital dis-

ruption and in the proportion of the sample black.

Between 57 and 100 percent of the increase in the receipt rate would not have been expected on the basis of these changes in the child support population. Contributing to this unexplained portion are significant changes in 4 out of 20 structural and 2 out of 19 reduced-form coefficients between 1978 and 1983. We attribute this to changes in attitudes toward nonpayment and in the enforcement of child support.

Among women receiving child support, real dollars received fell 15.6 percent (\$427) between 1978 and 1983. The reduced-form equation predicts receipts should have fallen only 3.2 percent (19 percent of the actual decline), while the structural equation, which includes the impact of a 19.4 percent drop in real dollars due, predicts that receipts should have fallen 17.2 percent (110 percent of the actual decline). Several factors other than the fall in real dollars due would also have suggested a decline in child support receipts in both the structural and reduced-form equations. In order of their contribution, these include the decline in average number of children, the increase in average time since marital disruption, the increase in the proportions of black and of never-married mothers receiving support, and the rise in the fraction of women residing in the South. Partly offsetting these changes were increases in the average age and educational attainment of women receiving support.

Since 81 percent of the decline in receipts in the reduced-form equation remains unexplained by changes in socioeconomic variables, one might be tempted to attribute this remainder to a decrease in social concern and less-effective child support enforcement. However, when the impact of the sharp decline in real dollars due is also taken into account, as in the structural equation, a very different conclusion emerges. What must have prevented an even greater decline in receipts was an increase in social concern and more effective enforcement. Differences between the structural and reduced forms also tell us something else. It is not possible to explain most of the drop in receipts unless we take account of the sharp decline in the

value of awards that occurred over this period of high inflation, due in large part to erosion in the value of nonindexed awards.

### V. Conclusion

In the past few years, many states have enacted new laws to enforce child support contracts, and government officials and the public at large have paid increasing attention to the problem of nonpayment. Unfortunately, published CPS data offer at best mixed evidence that these changes have had any significant impact: although the receipt rate rose, real dollars received fell between 1978 and 1983. Fortunately, our decomposition analysis offers more support on their behalf. It shows that more than half of the observed increase in the receipt rate would not have been expected on the basis of changes in the composition of the child support population alone over the period. Furthermore, given these socioeconomic changes and the sharp decline in real dollars of support due, we should have seen an even greater decline in dollars received than actually occurred. What prevented this greater decline in child support received, and what was responsible for the larger increase in its receipt

rate, were changes in the receipt process, evidenced by changes in our coefficients, which we attribute to improvements in the legal and social environment surrounding child support enforcement.

### REFERENCES

- Beller, Andrea H. and Graham, J. W., "Variations in Economic Well-Being of Divorced Women and Their Children: The Role of Child Support Income," in M. David and T. Smeeding, eds., *Horizontal Equity, Uncertainty, and Measures of Well-Being*, NBER Studies in Income and Wealth, No. 50, Chicago: University of Chicago Press, 1985, 471-509.
- \_\_\_\_\_ and \_\_\_\_\_, "Child Support Awards: Differentials and Trends by Race and Marital Status," *Demography*, May 1986, 23, 231-45.
- U.S. Bureau of the Census, *Current Population Reports*, Series P-23, No. 152 (148; 140; 112), *Child Support and Alimony: 1985*, (1983; 1981; 1978), Washington: USGPO, 1987 (1986; 1985; 1981).
- U.S. Office of Child Support Enforcement, *Child Support Enforcement*, 8th Annual Report, Washington: USGPO, 1983.



## Getting into Poverty Without a Husband, and Getting Out, With or Without

By THOMAS J. KNIESNER, MARJORIE B. MCELROY, AND STEVEN P. WILCOX\*

Interest in the poverty of U.S. women with children but without husbands stems from numerous sources including (i) the secular growth of this demographic group—up 110 percent since 1970 to a total of 6 million (almost 20 percent of all families) in 1985; (ii) the high poverty rates of these women—34 percent in 1985; (iii) the overrepresentation of blacks in this group—about 42 percent in 1985; (iv) the increasing fraction of children raised in these families—over 16 percent in 1984 vs. 6 percent in 1959; and (v) the size of government transfers to this particular group—almost \$17 billion for income support under the AFDC program alone in 1985.<sup>1</sup> Our research uncovers some important racial similarities as well as stark differences in how women enter and exit single-mother poverty status.

To address these concerns we analyze poverty spells of young single mothers. We pause to explain this break with the existing literature, that often examines spells of

poverty that transcend changes in family structure. Mary Jo Bane and David Ellwood (1986) pioneered this approach by tracking poverty spells for an individual across changes in family structure and classifying poverty spells according to the family structure at the start and, alternatively, at the end of the spell. This approach would, of course, not illuminate the origins of the class of women who are poor single mothers. Some of them would never show up as female household heads, either at entry into or exit from poverty. (For example, some poor single mothers started as poor and single; some poor single mothers go on to be poor and living with their parents.) Any poor single mother whose poverty began before her single-mother status did, or whose poverty terminated after her poor single-mother status did, would not ever be classified as a female head of household.

In light of the problems just mentioned, we focus directly on spells as a poor single mother, noting both family structure and poverty status immediately before and after. With some reluctance, throughout this paper we follow the Census Bureau's terminology and refer to a female head of household with dependent children as a female family householder (hereafter FFH) or, less formally, as a single mother.<sup>2</sup> For economy of language we therefore refer to spells of FFH

\*Department of Economics, University of North Carolina, Chapel Hill, NC 27599-3305; Department of Economics, Duke University, Durham, NC 27706; and General Research Co., McLean, VA 22105; respectively. We gratefully acknowledge support from the Small Grants program of the Institute for Research on Poverty, the University of Wisconsin, and the U.S. Department of Health and Human Services; the University Research Council and Economics Department of UNC-CH; and the NSF, Grant no. SES-8409262. Special thanks go to James Baumgardner, Sheldon Danziger, William A. Darity, Jr., David Ellwood, Carole A. Green, and Bob Hussey for many important insights; to Tony Maika for able research assistance; and to Sarah Mason, Forrest Smith, and Tom Kniesner for excellent typing. Stephen M. Hills provided invaluable technical assistance. Any opinions expressed here are solely our own and do not necessarily reflect the positions of any of the sponsoring institutions.

<sup>1</sup>The references for these data are (i) Bureau of the Census, *Statistical Abstract* (1987, p. 48); (ii; iii) *Statistical Abstract* (pp. 443 and 445); (iv) Victor Fuchs (1986, p. 12); and (v) *Statistical Abstract* (p. 343).

<sup>2</sup>A woman is defined as FFH if she is the head of her household and has children in the household. She is the head of her household if she is (i) *not* married with spouse present, and (ii) reports herself as either the head of the household or as the sister of head of household and no related adults (including the sister) are present. She is *not* the head of the household if she is either (a) married with spouse present or in the armed forces, or (b) does not meet (ii) above. By this definition, a divorced woman living with her parents would not be a FFH.

and spells of FFH poverty; similarly we use "divorce" to cover any marital split: divorce, separation, or absence of a spouse.

### I. Data and Spell Definitions

We use the *National Longitudinal Survey of Young Women (NLSYW)*, which oversamples blacks (28 percent) and follows a large number (5,159) of young women (ages 14 to 24 in 1967) for 14 years through stages of their lives when most marital and fertility transitions occur. Our research would not have been possible using either of the two standard data bases for poverty research: the *Current Population Survey*, the source of the official U.S. poverty rates, is basically cross sectional, and the *Panel Study of Income Dynamics* has too few women to focus on poor single mothers, let alone make meaningful comparisons by race.

We track the poverty status of each woman in the *NLSYW* for each survey date from 1968 to 1982 by following each family unit to which she belonged and comparing total family income to the relevant official annual poverty threshold.<sup>3</sup> Income here is the sum of each family member's income, *exclusive of* government in-kind and cash transfer payments.<sup>4</sup> The poverty thresholds are the offi-

cial ones established by the Census Bureau. Not only is official poverty status a simple index of economic well-being, but it is also often the basis for policy discussions concerning equality and the distribution of income. Because the official poverty thresholds are arbitrary meters of economic well-being, we checked the robustness of our results by varying the definition of poverty to 125 percent and 75 percent of the official thresholds by family type. All of the empirical results reported in subsequent sections are based on the official U.S. poverty threshold in the interest of brevity because those obtained with these two alternative poverty thresholds yield only two differences: raising the threshold accelerates the entry rate, slows the exit rate, and lengthens the duration of FFH poverty, and conversely; lowering the threshold increases the fraction of blacks who are FFH poor, as one expects, because the very poor are disproportionately black.

Finally, our analysis focuses on the first-observed FFH-poverty spell, which avoids oversampling from multispell individuals. Although there is information in second- and higher-order spells, they tend to be associated with second divorces or other repeated changes in family structure and are of secondary importance here.

### II. The Ins and Outs of FFH Poverty

The black rate of entry into FFH poverty is nearly quadruple the white rate of entry in our data; the mirror image of this is that the exit rate for blacks is only half that of whites. Thus, young black women are both much more likely to experience FFH poverty and, once in FFH poverty, have longer average spells, 3.8 vs. 2.5 years for the young white women in our sample. Data on year of exit by year of FFH-poverty entry for blacks and whites separately show that for most young

<sup>3</sup>The *NLSYW* did not survey the panel every year; 1974, 1976, 1979, and 1981 are missing. This gives rise to the following sort of problem. Suppose that a woman was FFH in 1973 and not FFH (hereafter denoted as NFFH) in 1975, with her status, of course, missing in 1974. Assuming, for the ease of discussion only, that she made one transition between the observed years, the question is was she out of FFH poverty in 1974, or not until 1975? We first explored the consequences of the two obvious possible extreme assumptions, the truth lying somewhere in between. At one extreme, we maintain that the spell ended in 1974. At the other extreme, we maintain that the poverty spell ended in 1975. We also made similar contrasting assumptions for the other missing years and spells of NFFH. Because our results were decidedly robust to these extremes, we report only the results for the first type of assumption—that the example spell ended in 1975 and that the other missing survey years were treated similarly.

<sup>4</sup>The income from a former (or absent) husband is also excluded. By ignoring cash transfers, the discrepancy between our poverty count and that of the federal government is the number of individuals who

are lifted out of poverty via such cash transfers. An appendix that contains more discussion of how we determine poverty status is available from the authors upon request.

women poverty is a short-term phenomenon. For a few, however, especially certain young black women, it is a long-term situation.

To discover what underlies the race differences in FFH poverty we have been discussing, we examined the entry and exit modes in the *NLSYW*. For both races, nearly all first observed FFH-poverty spells commence with a change in family structure; isolated drops in income mark the beginning of very few spells (0.6 percent for whites and 3.4 percent for blacks in our sample). However, the composition of changes in family structure differ markedly by race. Divorce is the prevalent entry mode for whites, but not for blacks: divorce accounts for 71.1 percent of white entries, but only 29.9 percent of black entries. Leaving the household of another adult is much more important for blacks, accounting for 45.8 percent of black entries and only 17 percent of white entries.

Most commonly, whites exit FFH poverty via remarriage 47.9 percent, while blacks exit via remarriage only 31 percent of the time. The dominant exit mode for blacks is what we have labeled "other change in family structure," the most typical one being rejoining their parent(s)' household or the household of an unrelated male adult. In contrast to FFH-poverty entry, isolated income changes play a significant role in FFH-poverty exit, terminating 33.3 percent of white and 29.9 percent of black spells.<sup>5</sup>

For each type of entry into FFH poverty we calculated the percentages who (i) were not poor in their prior family structure, (ii) were poor prior to poverty entry, or (iii) had unknown prior poverty status due to missing data. We also determined poverty status immediately following FFH-poverty spells. These cross tabulations clearly show that the bulk of FFH poverty is neither carried over from a woman's prior family status nor transmitted to her subsequent family status. For at least 75.1 percent of white and 51.4 percent of black entrants, FFH poverty is *new* poverty; similarly, at least 75.1 percent

of white and 58.7 percent of black exits are to nonpoverty.<sup>6</sup> For both races, those who enter FFH poverty via divorce are even more likely to be newly poor (at least 77.9 percent of whites and 66.2 percent of blacks). At the other end of FFH-poverty spells, of those that terminate via (re)marriage, at least 79.7 percent of white and 60.0 percent of black exits are to nonpoverty. Marriage clearly keeps young women of both races out of poverty, doing a somewhat better job for whites than for blacks.

In contrast, those blacks who enter FFH poverty via leaving the household of another adult or who exit FFH poverty via (re)joining the household of another adult were predominantly poor before entering (57 to 70 percent enter from poverty in another family structure) and after exit (49 to 74 percent exit to poverty in another family structure). The story here is different for the corresponding whites; they are mostly non-poor before FFH-poverty entry (at least 57.4 percent), but even *more* likely than blacks to exit to poverty (71.1 to 80.7 percent do). These racial differences in family structure and poverty patterns bring to mind the recent resurgent concern over the development of a permanent "underclass" within which poverty is passed from one generation to the next, especially in black families.

In closing this section it is important to note the findings of Bane and Ellwood, who use the *Panel Study of Income Dynamics* to summarize the poverty spells of all individuals (men, women, and children) whose spells began prior to age 64. They found an important role for drops in income for female heads' poverty spells: about 24 percent of their sample women who began a poverty spell as a female head did so via a reduction

<sup>5</sup>FFH-poverty spells that end because a child left account for less than 1 percent of the total.

<sup>6</sup>The minimum percentages and ranges of percentages appearing in this and the next paragraph reflect the fact that some women enter FFH poverty from unknown poverty status or exit FFH poverty to another family structure with unknown poverty status. We speculate that many of those with unknown poverty status in their original or destination family structures are poor.

in income. Superficially, this appears to contradict our finding of almost no role for income changes in entries. It needs to be emphasized that their sample of spells is, without a myriad of adjustments, noncomparable to ours for a number of reasons.

First, they used all spells of poverty, a procedure that, unlike ours, includes individuals' repeated spells. Because income changes occur more frequently than family structure changes, we speculate that their procedure emphasizes spells that begin or end with changes in income as opposed to changes in family structure. Second, Bane-Ellwood track women's spells of poverty through changes in family structure, rather than track spells of poverty in a particular family structure (FFH), as we do. Hence, their category of women who began a poverty spell as a female head would exclude all FFH poor women who began their current spell of poverty in some other family structure—while married, for example. Third, our women are exclusively young women (ages 14–24 at the beginning of the sample period), while theirs can start a poverty spell as late as age 64. Because our women are in the peak years for childbearing, divorce, and remarriage, this would account for some of the prominence that changing family structure has in our study vs. Bane-Ellwood. Fourth, we restrict our sample to young women heads with at least one dependent minor child. In contrast, their closest category is female headship. Thus, by not requiring the presence of a minor dependent child, they find that the earnings of adult children of the female heads play a major role both in their entry and exit from poverty. Finally, we examine pre-transfer income poverty, while they examine post-transfer poverty.

### III. Conclusion

Single mothers are a growing fraction of the U.S. population, are disproportionately black, often poor, raise an increasing fraction of U.S. children, and, along with their children, are the beneficiaries of the controversial U.S. welfare system. Our research is a first attempt to assess directly how young women become poor single mothers, de-

termine how they leave this situation, and ascertain how the events just mentioned differ between blacks and whites.

We find strong racial similarities in two broad and important areas. First, the relative importance of family structure and income changes in entering and exiting FFH poverty are the same for young black and young white women: changes in family structure account for nearly all entrances into FFH poverty; changes in family structure also dominate exits from FFH poverty. Isolated increases in income play an important secondary role, though. Second, to a surprising extent, for both races FFH poverty represents new poverty rather than poverty carried over from some previous family structure. Further, leaving FFH poverty means escaping poverty for most women, not merely experiencing poverty under a different family structure.

In contrast to the two racial similarities just noted, three crucial racial distinctions emerge from our research. The first is that blacks enter FFH poverty at much higher rates than do whites, exit at much slower rates, and hence average significantly longer spells of FFH poverty. The second race difference appears when we disaggregate changes in family structure: the dominant role of divorce and (re)marriage for whites and the parallel dominant role of leaving or (re)joining the household of another adult for blacks. Third, blacks who enter FFH poverty by leaving the household of another adult or who exit FFH poverty by joining the household of another adult have quite different poverty patterns than their white counterparts. These young black women are much more likely to be poor before and less likely to be poor after a spell of FFH poverty than the young white women who enter or leave FFH poverty through the routes just noted. This contrasts to those young women who enter FFH poverty via divorce or leave via (re)marriage. In this case, whites and blacks look similar; most are not poor before divorce and not poor after (re)marriage. The race differences in the link between poverty and family structure uncovered here highlight the need for much more research on the link between well-being and family

# Poverty Among Women and Children: What Accounts for the Change?

By LAURIE J. BASSI\*

Over the past few decades there has been a phenomenal increase in the number of households that are headed by women. This fundamental change in family structure has been accompanied by increases in the number of women in poverty, relative to the number of men in poverty. For instance, in 1959 women were 23 percent more likely to be poor than were men, but by 1985 they were 51 percent more likely to be poor.<sup>1</sup>

At the same time, the ratio of children's poverty, relative to that of men's has skyrocketed. While children were 53 percent more likely to be poor than were men in 1959, by 1985 they were 121 percent more likely to be poor. The "feminization of poverty" has become shorthand for describing the disproportionate percentage of poverty that is borne by women living alone or with their children.

This paper analyzes the forces behind these changes. A framework is developed that allows changes in poverty rates to be decomposed into component parts. Using a time-series of independent cross sections covering the period from 1967 to 1985, this framework is then used to examine the empirical significance of a variety of factors that contribute to poverty.

The first set of factors examined includes women's wage rates and hours of work, men's earnings, and AFDC-guarantee levels. The

effects that these variables have had on poverty rates are twofold; they have affected poverty directly (holding marital status constant), and have affected poverty indirectly through their effects on marital status. Both of these effects are considered. In addition, the direct (but not the indirect) poverty effects of changes in the number of children, and changes in income from child support and alimony, are examined.

## I. Empirical Framework

The poverty rate,  $PR$ , among women can be calculated from the following accounting identity:

$$(1) \quad PR = \sum_{i=1}^n \sum_{j=1}^2 PR_{ij} \cdot P_{ij},$$

where the subscript  $i$  represents the  $i$ th group of women,  $j$  denotes marital status, and  $P_{ij}$  is the percentage of women in the  $i$ th group who are in the  $j$ th marital status. Any factor that affects marital status will affect poverty rates among women and their children directly through its effect on  $PR$ , and indirectly through its effect on  $P$ .

Let  $X_k$  represent the  $k$ th such factor, and  $dPR_k$  represent the effect of a change in  $X_k$  on women's poverty. This effect can be written as

$$(2) \quad dPR_k = \sum_{i=1}^n \sum_{j=1}^2 \left[ \left( \partial PR_{ij} / \partial X_k \right) \cdot P_{ij} + PR_{ij} \cdot \left( \partial P_{ij} / \partial X_k \right) \right] \cdot dX_k.$$

Equation (2) can be rewritten in a simplified form which lends itself to estimation. The first simplification comes from recognizing that since there are only two marital states

\*Department of Economics, Georgetown University, Washington, D.C. 20057. I thank the Alfred P. Sloan Foundation and the Graduate School at Georgetown University for financial support. Ed Fu provided superb assistance in creating the data base used in the analysis, and Sheldon Danziger and Irv Garfinkel made useful comments on an earlier draft. I am particularly grateful to Shulamit Kahn for her patient and insightful discussions.

<sup>1</sup>Based on figures in Victor Fuchs (1986), and my tabulations from the CPS.

( $m$  = married and  $u$  = unmarried), it must be the case that  $\partial P_{im}/\partial X_k = -\partial P_{iu}/\partial X_k$ .<sup>2</sup> It is also true that  $PR_{im} = PR_{iu} + \partial PR_i/\partial M$ , where  $M$  is marital status. Finally, since the data that will be used in the analysis are grouped by marital status,  $P_{ij}$  is either equal to zero or one.

Using all of these identities, and some algebraic manipulation, equation (2) can be rewritten as

$$(3) \quad dPR_k = \sum_{i=1}^n [(\partial PR_i/\partial X_k) + (\partial P_{im}/\partial X_k) \cdot (\partial PR_i/\partial M)] \cdot dX_k.$$

The first term on the right-hand side of equation (3) represents how a variable directly affects poverty among women, holding their marital status constant. The second term represents how a variable affects women's poverty indirectly through affecting their marital status. Similar calculations can be made for children, where  $\partial P_{im}/\partial X$  in this case, represents how a variable affects their mother's marital status.

## II. The Data

In order to estimate each of the three derivatives given in equation (3), alternate years of the March Demographic files of the *Current Population Survey* from 1968 to 1986 were used to construct ten cross sections. Since the relevant questions from the CPS are asked retrospectively, 1967 to 1985 is the time period actually covered by the analysis.

Within each year the data were aggregated (using sample weights) by four age groups (18–24, 25–34, 35–44, and > 44), two race groups (black and white), standard metropolitan statistical area and state of residence, whether or not the individual lived in a central city, whether or not the individual

had a child under the age of 18, and marital status (never married, married, separated or divorced, and widowed). These groups then became the unit of observation. The earnings that unmarried women could expect to be generated by a spouse were determined by assigning to each group of unmarried women, the mean earnings of unmarried men of the same age, race, and geographic location.

The ten CPS cross sections were pooled for the purpose of calculating  $\partial P_{im}/\partial X_k$ . The estimates of these derivatives (taken from my 1987 paper, where these data were used, along with an explicit maximization model), are used to calculate the sources of marital status change over time. Both the intercept and the marital status response to (real) AFDC-guarantee levels were allowed to vary with time.<sup>3</sup> The other independent variables (all in real terms) included in the marital status analysis were women's wages and hours of work, and men's earnings. This implicitly constrains the derivative of marital status with respect to the number of children, and income from child support and alimony, to be zero.

The partial derivatives,  $\partial PR_i/\partial X_k$  and  $\partial PR_i/\partial M$ , that are needed as part of the calculations for equation (3), are taken from a regression where the dependent variable is the 1985 poverty rate for each group of women (or their children). The independent variables in the poverty regressions are: women's earnings, men's earnings, the AFDC-guarantee level in the state in which a group of women live, number of children under the age of 18, income from child support and alimony, and marital status.<sup>4</sup>

Finally, both the 1986 and 1970 files were used to calculate  $dX_k$ , the change in the independent variables between 1985 and 1969 (the year in which children's poverty reached its minimum). All regressions and calculations were done separately by age and

<sup>2</sup>It should be noted that the empirical analysis made distinctions between never married, separated or divorced, and widowed women. This allows differential marital status effects across these three different groups of unmarried women.

<sup>3</sup>In the empirical implementation, equation (3) was generalized to allow for these time-specific responses.

<sup>4</sup>Tables of the means of the variables, their changes over time, the regression coefficients and standard errors, are available upon request.

race group, and then aggregated (using sample weights), across age groups.

### III. The Results From The Analysis

Table 1 reports on the calculations that were generated from equation (3). These results can only be interpreted jointly with information about the direction of change in the exogenous variables (to be discussed below). It should also be noted that equation (3) uses results from two different sets of estimates (poverty regressions and marital status logits), to track changes over a 17-year period. Exercises such as this are fraught with many sources of forecasting error. The results substantially overestimate increases in black women's poverty (which was virtually unchanged between 1969 and 1985), and underestimate the increase in white children's poverty. Consequently, the estimates should be considered as only very rough approximations of the true orders of magnitude.

The results indicate that the increase in women's hours of work has been associated with an increase in poverty among women and children. The direct effect (holding marital status constant) is that an increase in women's labor market income reduces poverty among women and children. However, the indirect poverty effect resulting from the decrease in marriage that has been associated with (either as the cause of, or the result of) increasing labor market work among women, has more than offset these income increases.

Despite the fact that the number of female-headed households increased dramatically between 1969 and 1985, the level of child support and alimony was virtually constant, as was the level of women's wages. As a result, both variables had only a trivial effect on changes in poverty rates among women and children over this time period.

The effect of changes in men's earnings on poverty rates varies between whites and blacks. Over the time period under consideration, black men's earnings were increasing (with the exception of the youngest cohort), and white men's earnings were falling in real terms.

TABLE 1—SOURCES OF CHANGE IN WOMEN AND CHILDREN'S POVERTY RATES: 1969–85

	Black		White	
	<i>W</i>	<i>C</i>	<i>W</i>	<i>C</i>
Hours of Work by Mother	.025	.012	.027	.005
Mother's Wage	-.001	.000	.000	.000
Child Support and Alimony	-.001	.000	.001	-.001
Men's Earnings	-.006	-.014	.011	.018
Children under Age 18	-.034	-.068	-.005	-.014
AFDC Guarantee	.061	.039	.030	.033
Time <sup>a</sup>	.026	.049	-.059	-.048
Estimated Change	.070	.018	.005	-.007
Actual Change	-.005	.035	.003	.059

Notes: *W* stands for women, and *C* stands for children.

<sup>a</sup>Measures the effect of changes in the intercept in the marital status logit.

The earnings declines among white men have contributed to an increase in poverty among white women and children. This comes through two separate effects that reinforce one another. First, a decrease in men's earnings reduces the probability of marriage, thereby increasing the number of female-headed households (with disproportionately high poverty rates). At the same time, reductions in men's earnings increase poverty (of men, women, and children) within households that are headed by men.

The earnings gains among black men (with the exception of the youngest cohort) have contributed to a reduction of poverty among black women and children. Once again, both the marital status effect and the direct poverty effect contribute to a reduction in poverty. This finding is, however, sensitive to the choice of the base period. An earnings decline for black men started in 1973; so from 1973 to 1985, these decreases in black men's earnings contributed to an increase in poverty rates among black women and children.

The results in Table 1 indicate that decreases in family size (fewer children under the age 18) have generated fairly sizeable reductions in poverty rates, particularly among black women and children.

A striking result from Table 1 is the apparent importance of AFDC in determining

changes in poverty rates. These results indicate that between 1969 and 1985, the AFDC program was the most important source of increases in poverty among women and children (with the exception of black children for whom a pure time effect was the most important source of increasing poverty). This result is driven, in part, by a sizeable *decline* (35 percent) in real AFDC-guarantee levels, which has contributed to increasing poverty in female-headed households.

We would expect, however, that reductions in real AFDC-guarantee levels would increase marriage probabilities, and thereby tend to decrease poverty rates by reducing the number of households headed by women. However, the data seem to indicate that just the opposite has happened. Even though AFDC levels were falling, the marginal marital status response to it has increased.<sup>5</sup> The increasing marital status response to AFDC dominates the marital status effect of a decline in AFDC levels. The estimated net marital status effect is much larger for black women than for white women.<sup>6</sup>

There are, however, two important points to be made about this conclusion. The first is that, in general, white women had a slightly greater marginal marital status response to AFDC than did black women (both in 1969 and 1985). The *rate of increase* of marital status response to AFDC between 1969 and 1985 was, however, greater for black women than for white women; and it is the rate of change (rather than the level) of the response

that drives the result. The differential rate of change over time between blacks and whites may have been caused by systematic exclusion of blacks from receipt of AFDC throughout much of the 1960's (see Winifred Bell, 1965, for a discussion). With the lifting of the exclusions, blacks' response may have simply "caught up" (albeit incompletely) with whites' response.

The second important point to note about the apparent significance of AFDC in determining poverty rates, is that because of multicollinearity of the time dummies and the time-specific AFDC variables (in the underlying marital status logit), some of the coefficients used in the calculations were not statistically significant. This reduces the plausibility of the large estimated effect of AFDC on black women's poverty, particularly in light of the extent to which the change in their poverty is overestimated. While the underlying parameters track black women's marital status changes quite accurately, the analysis does a poor job of tracking changes in black women's poverty rates. The opposite is true for white women; changes in their poverty rates are forecast quite well, even though the analysis does not track changes in their marital status with a high degree of accuracy. (See my earlier paper for more detail on this point.)

Given these problems and the potential instabilities caused by multicollinearity, it seems clear that more work is needed before it is possible to convincingly disentangle the relative magnitudes of the effect of AFDC on poverty rates over time, from the effect of time itself. The estimates presented here do indicate, however, that *both* the welfare system and time itself have contributed to changes in poverty among women and children. The estimated effects of time itself, however, vary dramatically between whites and blacks. Surprisingly enough, the time-specific marital status intercepts indicate that the effect of the passage of time on white women's marital status has contributed to a *reduction* in poverty among white women and their children. The opposite appears to be true for black women. Why these time trends are so different between blacks and whites is unclear.

<sup>5</sup>See my earlier paper for a detailed discussion. There are two possible explanations for this phenomenon. The first is that there has been an increasing behavioral response to AFDC over time, perhaps due to a reduction in the social stigma associated with its receipt. The second is that AFDC merely serves as a proxy for the total welfare package available to unmarried women. Since AFDC has become a smaller percentage of the package, the marital status response to it may increase because of changes in the composition of the benefit package over time.

<sup>6</sup>In both cases, the indirect poverty effect of the increasing marital status response was substantially larger than the direct poverty effect of the reduction in real AFDC benefit levels.



#### IV. Conclusions and Policy Implications

The conclusions arrived at here are quite controversial, and at odds with those found by others (see, for instance, Thomas Kniesner et al., 1987). The most startling finding is that because the marital status response to the welfare system has been increasing over time, the system has become less effective in combating poverty among women and children. A variety of specifications of the underlying estimating equations were tested, and this result was found to be robust.

The relative importance of AFDC in determining poverty changes did, however, vary considerably for blacks and whites under alternative specifications. Nonetheless, under all specifications the effect of time (either through a changing marital status intercept or a changing marital response to AFDC) has had a much more profound effect on blacks than on whites.

Another striking finding is that increases in women's hours of labor market work have been associated with *increasing* poverty among women and children. While an increase in hours of work reduces poverty among existing female-headed households, it has at the same time been associated with (either as the cause of, or the result of) huge increases in the numbers of female-headed households.

The only concrete (but not surprising) policy conclusion that can be drawn from this analysis is that an effective method for combating increasing childhood poverty is to increase the earnings of low-income men. This has the immediate effect of reducing poverty among the children that these men support. At the same time, it increases the probability that these men will indeed support their children, since men become more "marriageable" with earnings gains.

Child support payments also have potential for reducing poverty among children. However, during the time period under consideration, child support payments were virtually constant. As a result, child support has done no more to relieve poverty among children in the 1980's than it did in the 1960's, despite the fact that there has been a dramatic increase in the number of children in single-parent households.

There is much left to be learned about the causal forces behind the "feminization of poverty." The methodology used here of pooling time-series and cross-section data appears to be a useful one. It allows for identification of how the economic determinants of marital status, and therefore of poverty rates, have changed over time. In the future, research should focus on identifying time-specific responses, rather than assuming that these responses are constant over time. In addition, a better understanding of the differences between blacks' and whites' marital status patterns is needed.

#### REFERENCES

- Bassi, Laurie J., "Changing Family Structure: Economic Choice or Economic Constraint?," mimeo., Georgetown University, October 1987.
- Bell, Winifred, *Aid For Dependent Children*, New York: Columbia University Press, 1965.
- Fuchs, Victor R., "The Feminization of Poverty," NBER Working Paper No. 1934, 1986.
- Kniesner, Thomas J., McElroy, Marjorie B. and Wilcox, Steven P., "Family Structure, Race, and the Hazards of Young Women in Poverty," mimeo., University of North Carolina-Chapel Hill, July 1987.

## MARKETS FOR INFORMATION<sup>†</sup>

### Selling and Trading on Information in Financial Markets

By ANAT R. ADMATI AND PAUL PFLEIDERER\*

It is evident that markets for information, taking a variety of forms, are an important element in financial markets. Information-related commodities include newsletters, security analysis, active portfolio and fund management, and investment advisory services. There is a host of important and interesting issues that these information markets raise. Among them, we discuss here whether an information owner will wish to sell information where the alternative is to trade strategically on the basis of the information. If selling is desirable, we discuss the selling method that is optimal for the information seller. In addition to obtaining some insights on how information markets operate, this research sheds light on the equilibrium allocation of information in financial markets.

Suppose that one agent in a financial market has private information that other agents do not have. For most of this paper we will examine the possibility that this agent engages in the direct sale of his information to other traders in the financial market. In a direct sale, the buyer of the information actually observes the information and then uses it as he or she wishes for trading. We will allow the information owner to trade on his own account on the basis of the information, as well as sell information. In order to focus the discussion and to allow informa-

tion markets to operate whenever this is desirable, incentive problems will not be dealt with—it will be assumed that the statistical properties of the information are commonly known and that if information is sold it is communicated truthfully.

We first ask how the optimal selling and trading strategy depends on the risk tolerance of the traders and on the precision of the information, and show how the total profits of the informed trader change as these parameters vary. We find that if the information owner is risk neutral, he does not wish to sell his information to any other trader. This is due to the fact that informed traders compete with each other, reducing their total profits. If the seller is risk averse, however, then it is generally profitable for him to sell the information to other traders. Despite the competition that information sales create among informed traders, they allow better risk sharing, which is desirable. We will examine the allocation of information that is optimal for the information owner and show that, under some conditions, the information owner would actually like to commit not to trade personally on the basis of the information.

Another way to “sell” information is for the information owner to create a mutual fund (see our 1987a paper). Investors in a mutual fund do not observe the information directly; instead they purchase shares in the fund and the fund uses the information to determine its trading position. By charging each investor a fee that is a function of the number of shares the investor buys, the information owner effectively charges for the response to the information. By setting the fees appropriately, the information owner can control the effects of competition among these indirectly informed traders and increase his profits.

<sup>†</sup>*Discussants:* Douglas W. Diamond, University of Chicago and Yale University; Boyan Jovanovic, New York University; Richard Rogerson, University of Rochester and Southern Illinois University.

\*Graduate School of Business, Stanford University, Stanford, CA 94305. We thank Doug Diamond and David Kreps for comments on an earlier draft. Financial support from the Robert M. and Anne T. Bass Fellowship, Batterymarch Financial Management, the Sloan Foundation, and the Stanford Program in Finance is gratefully acknowledged.

Our analysis will be performed in the context of a specific model, based on Albert Kyle (1984, 1985), in which traders submit market orders and take into account their effect on the price. The results will be related to previous literature, particularly to our previous work (1986, 1987a), where we considered markets for information that is used in a perfectly competitive noisy rational expectations equilibrium model. In our previous work we assumed that the information owner does not trade on the basis of the information, or that if he does trade, he behaves competitively and ignores his effect on the price. Then we examined the optimal way to sell information. Since the information owner is "large" in the sense that his profits have the same order of magnitude as the total profits of all traders, it seems appropriate to consider the possibility, discussed here, that in fact he is strategic in trading on the information, taking into account his effect on the price.

### I. The Model

There are three time periods. In period 0, information may be exchanged, thereby determining the allocation of information. In period 1, trading takes place in the financial market. Finally, payoffs are realized and consumption takes place in period 2. The financial market is best thought of as a futures market. The spot price, realized in period 2, is a random variable  $\tilde{q}$  with mean  $\bar{q}$  and variance 1. (The normalization of  $\text{var}(\tilde{q})$  to 1 is without loss of generality.) In addition to the futures contract, a riskless asset is available, whose price and payoff are normalized to 1. There is a large number of potential speculators, each of whom maximizes his expected utility of final consumption given the information he or she possesses. In addition, there are some hedgers, whose total position in the futures market is random and independent of the spot price and of any private information.

We assume that there is one agent who has private information concerning the spot price  $\tilde{q}$ . This agent observes the signal  $\tilde{q} + \tilde{\theta}$ , where  $\tilde{\theta}$  and  $\tilde{q}$  are independently distributed, and may sell it to some of a large

number of potential traders. In a direct sale of the information, the seller announces the price of the signal (before the signal is realized), and a trader who decides to buy the information pays the price and then gets to observe the signal. This occurs prior to trading in the futures market, so that the trader can use the information in making his trading decision. As already mentioned, we assume that incentive problems do not arise whether or not the information owner trades on his own account. It is also assumed that information cannot be resold. (This may be justified by thinking of the information as "short lived" in the sense that its value declines quickly, so that it is basically infeasible to set up a secondary market in the information.)

The asset market model is based on Kyle (1984, 1985), with many potential traders (who may be risk averse) and a competitive, risk-neutral, market maker. Each trader submits a market order to a market maker. The market maker sets a price that is a function of the total (net) order flow so that his expected profit conditional on the total order flow is zero. For example, suppose that there are  $n$  informed traders, each of whom observes the signal  $\tilde{q} + \tilde{\theta}$  and submits the market order  $\beta(\tilde{q} + \tilde{\theta})$  for some constant  $\beta$ . Denote the total amount of hedging demand by  $\tilde{z}$ . Then the total market order is  $n\beta(\tilde{q} + \tilde{\theta}) + \tilde{z}$ , and the price is given by  $\tilde{p} = E(\tilde{q} | n\beta(\tilde{q} + \tilde{\theta}) + \tilde{z})$ . Note that the order flow provides information to the market maker, since the informed component of the order flow is correlated with the spot price. Each trader submits his market order realizing his effect on the total order flow and therefore on the price. The equilibrium in the futures market is a Nash equilibrium among the speculators.

An important observation is that uninformed traders who are risk averse will not wish to buy or sell any positive amount of the futures contract. This follows because the futures price is set equal to the conditional expected spot price given information (namely the total order flow) that is finer than what the uninformed have. Thus, any portfolio which includes the futures contract is stochastically dominated by a portfolio

which involves only the riskless asset. The number of active traders in the futures market is therefore equal to the number of informed traders and hedgers. Even if uninformed traders are risk neutral, there is no equilibrium in which they invest nonzero amounts in the futures market, although the argument is more complex.

The value of information to an individual trader is defined as the amount of money that equates the trader's *ex ante* expected utility with and without the information. (This calculation is performed prior to the actual observation of the information.) The profits of an informed trader in this market clearly depend on the total number of informed traders—the more informed traders there are, the lower are the trading profits and the lower is the value of information to each one of them. This is due to the competition among informed traders. As a result, for each price that the information seller sets for the information, there is a equilibrium number of traders who would buy the information at that price. These calculations are discussed in greater detail below.

For tractability we will assume that all random variables are normally distributed and that traders' preferences exhibit constant absolute risk aversion, that is, their utility functions are negative exponential.

## II. Direct Sale of Information

Suppose first that all potential traders, including the information owner, are identical. It is clear that, in this case, the information owner will always trade on his information, as the value of the information to him as a trader is identical to that of any other trader. The question is whether he would also like to sell the information to other traders. Since the seller is a monopolist who can extract all consumers' surplus, his profits are always equal to the total certainty equivalent of the trading profits of all the informed traders from trading in the futures market. This follows because, as discussed above, an uninformed agent does not trade in the futures market.

To determine the profits of the information owner as a function of the number of

informed traders we must derive the equilibrium in the asset market and calculate the value of information for each number of informed traders. Suppose that there are  $n$  informed traders, and denote the total order flow by  $\tilde{\omega}$ . Suppose that the market order of each informed trader is a linear function of his information. Then  $\tilde{\omega}$  is a normal random variable, and the pricing function can be written as  $\tilde{p} = \bar{q} + \lambda \tilde{\omega}$  for a constant  $\lambda$ .

To find the equilibrium, we first derive the reaction function of each trader to the strategies of others, taking the pricing function as given. Not surprisingly, the unique equilibrium with identical traders is symmetric. Suppose each of  $n-1$  informed trader submits the market order  $\beta(\tilde{q} + \tilde{\theta})$  for some constant  $\beta$ . Then the second period wealth of the  $n$ th informed trader if his market order is  $x$  is given by

$$(1) \quad W_0 + x(\tilde{q} - \tilde{p}) = W_0 + x(\tilde{q} - \bar{q} - \lambda(x + (n-1)(\tilde{q} + \tilde{\theta}) + \tilde{z})),$$

where  $W_0$  is the initial wealth of the trader. Using the assumptions of exponential utility and normal distributions, the objective function of the trader is equivalent to choosing  $x$  to maximize

$$(2) \quad E(x(\tilde{q} - \tilde{p})|\tilde{q} + \tilde{\theta}) - \frac{a}{2} \text{var}(x(\tilde{q} - \tilde{p})|\tilde{q} + \tilde{\theta}),$$

where  $a$  denotes the coefficient of risk aversion. It is straightforward to derive the reaction functions and solve for the symmetric Nash equilibrium among informed traders. For a random variable  $\tilde{x}$ , denote  $\text{var}(\tilde{x})$  by  $\sigma_x^2$ . Then for a given price coefficient  $\lambda$  the equilibrium value of  $\beta$  is given by

$$(3) \quad \beta = (a\lambda^2(1 + \sigma_\theta^2)\sigma_z^2 + (n+1)(1 + \sigma_\theta^2)\lambda + a\sigma_\theta^2)^{-1}.$$

Now, given these strategies by the informed traders, the market maker has to choose  $\lambda$  to satisfy the zero expected profits

condition. From the theory of normal variables and the assumptions of the model it follows that

$$(4) \quad \lambda = \frac{\text{cov}(\tilde{q}, \tilde{\omega})}{\text{var}(\tilde{\omega})} = \frac{n\beta}{n^2\beta^2(1 + \sigma_\theta^2) + \sigma_z^2}.$$

Substituting the value of  $\beta$  from (3) in (4) and solving the resulting equation yields the unique symmetric equilibrium.

Suppose first that all traders are risk neutral (i.e.,  $a = 0$ ). Then the above expressions simplify greatly and the linear equilibrium coefficients are solutions to linear equations. Moreover, with risk neutrality, the value of information is simply the *ex ante* expected profits of each trader in equilibrium.

Denote the value of information to each trader by  $\pi(n)$  when there are  $n$  informed traders in the market (including possibly the information owner). It is straightforward to show that *with risk-neutral traders, the total profits of the informed traders,  $n\pi(n)$ , are decreasing in  $n$* . This is analogous to the result that industry profits are decreasing in the number of firms in a Cournot oligopoly model. The intuition is that competition among informed traders leads to lack of coordination and therefore reduced industry profits. Since the owner of information in our model can determine the number of informed traders in the market, this result implies that *under risk neutrality, information will not be sold*. That is, the information owner prefers to trade as a monopolistic trader rather than to sell the information to other potential traders and extract their trading profits.

It is easy to see that the above argument can be generalized to show that information will not be sold in our model if the information owner is risk neutral, as long as other traders are not risk seeking (i.e., independent of other traders' preferences).

The situation changes if the information owner is risk averse. Let us start again by assuming that all traders have the same preferences with common coefficient of risk aversion  $a > 0$ . The model with risk aversion is analytically complicated, since the linear equilibrium involves the solution to a fifth-

order equation. Note that to find the value of information we must first calculate the *ex ante* expected utility of an informed trader. Let  $\phi = \text{var}(\beta(\tilde{q} + \tilde{\theta})) = \beta^2(1 + \sigma_\theta^2)$ ,  $\psi = \text{var}(\tilde{q} - \tilde{p}) = \beta(1 - \lambda n\beta(1 + \sigma_\theta^2))$ , and  $\gamma = \text{cov}(\beta(\tilde{q} + \tilde{\theta}), \tilde{q} - \tilde{p}) = 1 - n\beta\lambda$ . Then the *ex ante* expected utility of an informed trader in this market (calculated before the information is observed) is given by

$$(5) \quad -((a\gamma + 1)^2 - a^2\phi\psi)^{-1/2} \\ \times \exp(-a(W_0 - c)),$$

where  $c$  is the price of the information. It follows that the value of information, that is, the maximum that a trader would be willing to pay to become informed, is

$$(6) \quad \frac{1}{2a} \log((a\gamma + 1)^2 - a^2\phi\psi).$$

It is possible to find the equilibrium values of  $\lambda$  and  $\beta$  for different values of  $n$  by numerical methods. Then the above equations can be used to calculate the value of information  $\pi(n)$  as a function of the number of informed traders. As before,  $n\pi(n)$  is the total amount that the informed traders are willing to pay in order to participate in this market, and it is equal to the profit of the information owner when there are  $n$  informed traders (including himself).

We find that, unlike the case of risk-neutral traders, *if traders are risk averse then these total profits may be increasing in the number of informed traders  $n$* . For example, if  $\sigma_\theta^2 = \sigma_z^2 = a = 1$ , then  $\pi(1) = .17285$ , while  $4\pi(4) = .22341$ . In this case, in fact, having four informed traders is optimal for the information owner.

We have seen that if the information owner is risk averse, he may want to sell his information to other traders in addition to trading on it. Intuitively, in some cases risk sharing is sufficiently valuable to compensate for the losses caused by competition. Generically, the optimal number of informed traders is unique, since  $n\pi(n)$  is a unimodal function, increasing for small  $n$  and decreasing

for large  $n$ . That is, when  $n$  is small, the effect of risk sharing is stronger, while when  $n$  is large, the competition effect is stronger.

The number of informed traders that maximizes the profits of the information owner is (weakly) increasing in the coefficient of risk aversion. This is intuitively clear, since risk sharing is more valuable when traders are more risk averse. Not surprisingly, the total profits of the information owner are lower if the market participants are more risk averse. In this case, information is generally less valuable and individuals act less aggressively in response to information. (For a proof of this result in the context of a perfectly competitive rational expectations model, see Robert Verrecchia, 1982.) At the other extreme, we have seen above that when agents are risk neutral, the optimal number of informed traders is one, so that information is not sold at all.

It is also interesting to examine how the number of informed traders and the information owner's profits change as the precision of the information increases (i.e., as the variance of the error term  $\tilde{\theta}$  in the signal declines). It turns out that less precise information is sold to more traders, and leads to lower profits for its owner. The result that the information owner's profits are increasing in the precision of his information is in sharp contrast to our results (1986), where we showed that an information seller may want to add noise to his information before selling it. (It was also shown there that if the optimal type of noise is added then the seller always sells identically distributed information to all the traders.) In our earlier paper, traders determine demand functions and can therefore free ride on the information held by others. This is because private information is aggregated in the price and, in the rational expectations model, traders can use the price to make their investment decision. In the model analyzed here, traders do not know the price before they submit their order, and hence they cannot use it in making their decisions. There is no leakage of the information from those who buy it to others who don't.

To conclude this section we examine the trading and selling strategy of the informa-

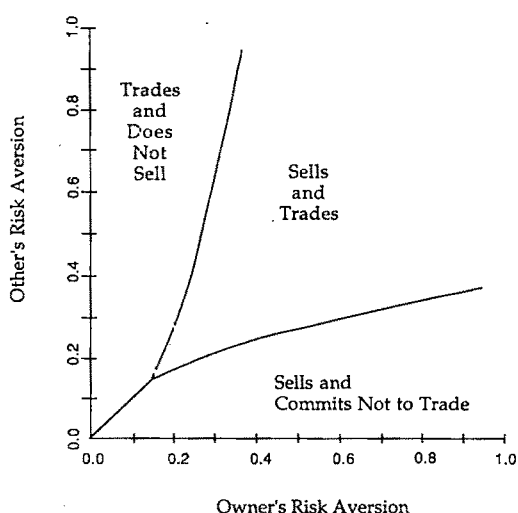


FIGURE 1

tion owner when agents have differing levels of risk aversion. We have already seen that if the information owner is risk neutral, then he will not sell information to any other potential trader. On the other hand, if the information owner is risk averse and there is another potential trader who is risk neutral, then it is clear that the information owner prefers to sell the information to the risk-neutral trader and commit not to trade himself on the basis of the information. (If a commitment not to trade is impossible, then it may still be optimal to sell the information, although if the information owner is sufficiently risk tolerant, he would prefer not to sell.)

If the information owner and all other potential traders are risk averse, then the optimal allocation of information depends on the absolute and the relative magnitudes of the risk-aversion coefficients. To illustrate what happens in this case, we consider a simple example where there is only one potential speculative trader in addition to the information owner. In Figure 1 we show, for the case  $\sigma_\theta^2 = \sigma_\epsilon^2 = 1$ , the allocation of information that is optimal for the owner for each pair of risk-aversion coefficients. For example, if the owner's coefficient is .15 and the other trader is relatively risk averse (precisely, if his risk-aversion coefficient is above

.16), then it is optimal for the owner to trade on the information and not sell it. If, instead, the other trader's risk-aversion coefficient is below .1456, then it is optimal for the information owner to sell the information and to commit not to trade himself. For intermediate range of risk-aversion coefficients for the other trader, it is optimal for the owner to sell the information as well as trade on it. Note that if either the information owner or the other trader is sufficiently risk tolerant (if either coefficient of risk aversion is below .1421), then it is never optimal to have more than one informed trader. In this case the more risk-tolerant trader is the only one informed.

The intuition behind the above results is straightforward given our earlier discussion. There is a tradeoff between coordination (i.e., less competition) and risk sharing, where more competition lowers the value of information and risk sharing increases it. This is combined with the fact that the more risk tolerant a trader is, the more aggressive he is in using information, and the higher is the value of information to him, other things equal. This latter fact means that it cannot be optimal for information to be held by a risk-averse trader if there is a more risk-tolerant potential trader who is uninformed. In other words, the allocation of information must be *viable* in the sense of our earlier article (1987b).

When there are many potential traders with different preferences, price discrimination becomes an important issue, since the value of information is decreasing in the risk-aversion coefficient. If the information owner sells information to traders with different preferences and price discrimination is impossible, then he will lose some of the consumers' surplus.

### III. Indirect Information Sales through Funds

Here we examine the possibility that the information owner sells information *indirectly* through a mutual fund. This takes the following form. For each share in the fund, the owner submits a market order of  $\tilde{q} + \tilde{\theta}$ . Each trader who buys shares in the fund is charged a fixed fee and a per share

fee. With identical traders, this is the most general pricing scheme. Note that by setting the per share price to zero and charging the appropriate fixed fee, the information owner can obtain the same profits as in a direct sale of information.

We assume first that the information owner commits not to trade in the futures market. If all potential traders have identical preferences, then it is clear that in equilibrium all of those who purchase shares in the fund purchase the same number of shares. Let each of  $m$  traders buy  $x$  shares. Then the fund's total market order is  $mx(\tilde{q} + \tilde{\theta})$ , and each investor receives the random amount  $x(\tilde{q} + \tilde{\theta})(\tilde{q} - \tilde{p})$  in period 2.

When selling shares in the mutual fund, the information owner is a monopolist facing an aggregate demand function. It is important to note that even when all potential traders are identical, the aggregate demand function is not the sum of the individual demand functions. To see this, recall that the value of information depends on the total position of all traders, which in this case depends on the total number of shares of the fund. The value an individual investor places on an additional share in the fund is different depending on whether the position of other investors remains constant or changes.

Denote by  $v(x, X)$  the value an investor places on holding  $x$  shares in the fund when the total number of shares purchased (including his own) is  $X$ . By adjusting the per share fee, the information owner can induce investors to choose any number of shares in a range of values. By adjusting the fixed fee, he can extract the consumers' surplus, as well as determine the number of traders who buy shares in the fund. The objective function of the information owner is then to maximize  $mv(x, mx)$  over  $m$  and  $x$ .

Recall that in the symmetric equilibrium of the model in which information is sold directly, each informed trader submits a market order of  $\beta$  times the signal  $\tilde{q} + \tilde{\theta}$ . Let  $\beta(m, a)$  be the equilibrium value of  $\beta$  when there are  $m$  informed traders, all of whom have the same coefficient  $a$  of risk aversion. Turning to the case of a mutual fund, it can be shown that for a given  $m$ , the optimal value of  $x$  is  $\beta(1, a/m)/m$ . That is, the

mutual fund trades as if it were a single trader having risk aversion equal to  $a/m$ . (In effect, the fund is equivalent to a syndicate as in Robert Wilson, 1968.) It can also be shown that the owner's profits are increasing in  $m$ , so the owner induces all potential traders to buy shares. (If potential traders have different preferences and price discrimination is impossible, then the information owner may set prices so that not all potential traders purchase shares in the fund.)

To obtain some intuition, note that for all  $m > 1$ ,  $m\beta(m, a) > \beta(1, a/m)$ . That is, with direct sale of information the aggregate position of the  $m$  traders is larger than the fund's position. The effect can be substantial: if  $\sigma_x^2 = \sigma_\theta^2 = a = 1$ , and information is sold directly to 100 traders, then in equilibrium each trader submits a market order of  $.03299(\tilde{q} + \tilde{\theta})$  and each has a certainty equivalent for the trading profits equal to  $.0008815$ . If instead 100 traders buy shares in a mutual fund, then the optimal total position for the fund is  $.70094(\tilde{q} + \tilde{\theta})$ . This is significantly lower than  $3.2986(\tilde{q} + \tilde{\theta})$ , the total market order placed by the 100 traders when they buy information directly. Thus, the mutual fund restrains the total trading response to the signal.

In the example above, the value of each trader's position in the mutual fund is  $.0034925$  or roughly four times the value of the information when it is sold directly. The owner's profits from selling shares in the fund are therefore  $.34925$ . As we saw in Section II, the highest profits of the information owner when selling information directly are attained when there are exactly four informed traders, and total profits in this case are  $.22341$ . Thus, there is a gain to selling information through the fund.

The above analysis was performed under the assumption that the information owner does not trade in the futures market on his own account. Now suppose that the information owner can commit to hold a specific number of shares in the fund, and can also commit not to trade outside the fund. If the owner's preferences are identical to those of all potential traders, then it is clearly optimal for the owner to commit to holding the

same number of shares as all other traders hold in equilibrium. In this case, the above analysis applies, where the set of potential traders includes the information owner. If commitment is impossible, then buyers of shares will take into account the fact that the information owner will compete against the fund in the futures market, and this will lower the owner's total profits. However, profits will still be higher than they would be if information is sold directly.

#### IV. Concluding Remarks

We have seen that when an information owner is risk averse, it becomes desirable to sell information. The exchange of information involves a number of incentive problems that we have not considered here. These problems are particularly important if the information owner both sells information and trades on his own account, as we have allowed above.

We have also seen that with identical traders and an ability to sell shares in a mutual fund using a two part pricing scheme, selling information indirectly is better than selling it directly. However, we do observe the direct sale of information (via newsletters and advising services). As we have argued earlier (1987a), direct sale may dominate indirect sale when there are many financial assets and many sources of information. For example, if potential information buyers are also endowed with different private signals, they may not find it desirable to invest through a fund, since a fund is a "bundled" response to a particular set of signals. Direct information sale, on the other hand, allows the trader to unbundle the different components of the information and combine it optimally with his other pieces of information.

#### REFERENCES

- Admati, Anat R. and Pfleiderer, Paul, "A Monopolistic Market for Information," *Journal of Economic Theory*, August 1986, 39, 400-38.  
 \_\_\_\_\_ and \_\_\_\_\_, (1987a) "Direct and In-



- direct Sale of Information," Research Paper No. 899, Graduate School of Business, Stanford University, September 1987.
- \_\_\_\_ and \_\_\_\_\_, (1987b) "Viable Allocations of Information in Financial Markets," *Journal of Economic Theory*, October 1987, 43, 76-115.
- Kyle, Albert S., "Market Structure, Information, Futures Markets, and Price Formation," in Gary G. Storey et al., eds., *International Agricultural Trade: Advanced Readings in Price Formation, Market Structure, and Price Instability*, Boulder, London: Westview Press, 1984, 45-64.
- \_\_\_\_, "Continuous Auctions and Insider Trading," *Econometrica*, November 1985, 53, 1315-35.
- Verrecchia, Robert E., "Information Acquisition in a Noisy Rational Expectations Economy," *Econometrica*, November 1982, 50, 1415-30.
- Wilson, Robert, "On the Theory of Syndicates," *Econometrica*, January 1968, 36, 119-32.

# Informational Theories of Employment

By BETH ALLEN AND COSTAS AZARIADIS\*

We report in this paper some elementary consequences of defining commodities not only by their physical properties, date of delivery, and state of nature—as one normally would in competitive equilibrium theory—but also by an informational attribute called “observability.” The coordination of trading plans in an economy with unobservable goods requires that each household should satisfy the standard budget constraint for observable goods plus additional constraints for unobservable goods that may be otherwise physically identical to observable goods. These extra market-by-market restrictions induce competitive equilibria that neither exhaust all gains from trade nor necessarily achieve constrained Pareto efficiency. As an application, we explore intertemporal labor supply by workers with unobservable leisure endowments. We provide an example of deficient aggregate employment, derive a sufficient condition for the constrained inefficiency of all *laissez-faire* equilibria, and illustrate how a relatively uninformed government might Pareto-improve equilibrium allocations by the judicious use of national debt.

Indexing commodities by informational attributes seems to be a direct and parsimonious way of rationalizing a wealth of empirical observations on aggregate time-series as well as many long-standing features of the labor and credit markets. Among these phenomena we count disposable income as an explanatory variable for aggregate consumption (see the survey by Robert Hall, 1987), corporate profits as an explanatory variable for aggregate investment, and the ubiquity of nonprice rationing for blue-collar workers (see the monograph by Russell

Cooper, 1987) and borrowers (compare Glenn Hubbard and Kenneth Judd, 1986). The equilibrium behavior of aggregate quantity variables like consumption, investment, employment, and lending is not easily reconciled with the standard model of general equilibrium (as reported in Gerard Debreu, 1959) if we continue to regard demand schedules as private responses to *observable* price signals in competitive markets; we propose instead that demand schedules are in part responses to *unobservable* shadow prices.

Unlike market prices, shadow prices are specific to individuals and depend on the tightness of quantity “rations” each person may face. How do such rations arise? And who sets them? Rationing in its most general form is a nonlinear price schedule relating the unit price of a given commodity to the amount purchased by a particular individual. Nonlinear prices are often thought to be outcomes of contracts between two agents (or of more complex allocation mechanisms among several agents) under conditions of asymmetric information—for example, moral hazard or adverse selection (see the survey by Oliver Hart and Bengt Holmstrom, 1986). To implement these contracts, traders rely on some publicly observed signal (often net sales or purchases) that is correlated with the relevant privately observed variable.

What happens if all publicly available signals are weakly correlated with the relevant private observation or, in extreme cases, completely orthogonal to it? We describe below (Section I) how this shortage of informative signals results in market-by-market budget constraints and eliminates trades that would be feasible in more informative environments; an example illustrates (Section II) how the price of a commodity may reflect its “observability.” We then study (Section III) the intertemporal equilibria of an economy with unobservable dated commodities, explore the effect of credit rationing on aggregate labor supply, and the circumstances

\*Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297. We thank Bernard Dumas for useful comments and the National Science Foundation for research support.

under which government intervention may be beneficial (Section IV).

# I

Consider first a pure exchange economy with  $L$  consumption goods indexed  $l = 1, \dots, L$  by their physical characteristics, delivery date, state of nature, etc. There are  $H$  households or agents indexed  $h = 1, \dots, H$ , each of them endowed with deterministic stocks of at least one good and with a deterministic preference ordering over all goods. Agents do not know each other's utility functions but may be able to observe some of the net trades of others. A given financial intermediary, for instance, may know the excess demand for loans by some individuals (its own customers) but not those of other individuals (the customers of other intermediaries). An employer may observe how much labor is supplied by each of his or her own employees but be completely ignorant of labor supplied elsewhere; finally, individual holdings of physical capital or of durable goods can be ascertained more easily than endowments of human capital.

To define more precisely the distribution of information over individuals, we denote by  $I = \{1, \dots, L\}$  the set of all households and by  $I_{lh} \in I$  the set of all agents who observe the net trade of goods  $l$  by household  $h$ . We suppose that an agent is either fully informed or completely in the dark (but never partly informed) about the net trade or position of each household in each market. If there are more than two goods, for example, Peter may know Paul's trades in bananas and Mary's position in apples without knowing Paul's actions in the apple market or Mary's in the banana market.

It makes sense for people to be aware of their own actions, which means  $h \in I_{lh}$  for all  $(h, l)$ . Apart from that, our framework will accommodate any information structure. An agent's trades of commodity  $l$  are *public knowledge* or *verifiable* if they are observed by all (i.e., if  $I_{lh} = I$ ). They are *private knowledge* if they are observed only by the individual in question (i.e., if  $I_{lh} = \{h\}$ ). We say that commodity  $l$  is *observable* if  $I_{lh} = I$  for all  $h$ , *unobservable* if  $I_{lh} = \{h\}$  for all  $h$ ,

and *partially observable* if it is neither "observable" nor "unobservable." Some of the better-informed participants in this economy may be able to contract with or intermediate between others, especially if the net trades of informed individuals are public knowledge. For example, agents who observe the outcome (or signals strongly correlated with it) of entrepreneurial investment projects can intermediate between the investor and ultimate lenders, writing a loan contract with the former and deposit contracts with the latter.

Contracts, however, are not our main concern. In order to keep the exposition focused on competitive trading processes, we rule out informed intermediaries and related forms of differential information. To that end, we suppose that for each  $l$  and  $h$  either  $I_{lh} = I$  or  $I_{lh} = \{h\}$ ; in words, net trades of any good by any agent are either fully "public knowledge" or strictly "private knowledge." This does not mean that every physical good is either observable or unobservable: Peter's apple trades could be known to all while Paul's may only be known to himself.

To understand how budget constraints coordinate trading in an economy with unobservable goods, it is useful to pretend that agents cannot trade unless they can establish ownership of the commodities they offer for sale. Accordingly, we imagine hypothetical "certificates of ownership" available freely to anyone whose holdings of a particular commodity are public knowledge, and not available at any cost otherwise. For each agent  $h$ , we define now the sets  $O_h$  and  $U_h$  of all commodities whose net trades by  $h$  are public knowledge and private knowledge, respectively. We express budget constraints in the following way:

$$(1) \quad \sum_{l \in S_h} p_l z_l^h + \sum_{l \in O_h} p_l z_l^h \leq 0$$

for all  $S_h \subset U_h$ .

Here  $z_l^h = x_l^h - \omega_l^h$  is the excess demand of agent  $h = 1, \dots, H$  for commodity  $l = 1, \dots, L$ , i.e., the difference between the agent's consumption and endowment of that commodity. Observe that we obtain one

budget constraint for each subset  $S_h$  of the set  $U_h$  of unobservable goods. For example, if  $S_h$  is empty, expression (1) reduces to the standard budget constraint over observable commodities only. If, on the other hand,  $S_h = U_h$ , then we obtain the standard budget constraint over all goods—unobservable as well as observable. These constraints prevent individual  $h$  from financing shortages of observed goods with surpluses of unobserved ones because the ownership of unobservables cannot be established. On the other hand, agents can pay for unobserved commodities by supplying observed ones.

## II

One competitive process we can study in this framework is the inside of an Edgeworth box. Suppose, then, that an economy consists of two goods and two (types of) agents. Net trades of good 1 (the numeraire commodity) are observed by all, net trades of good 2 are not. Letting  $p$  be the price of good 2, we find easily that inequality (1) reduces to the standard budget constraint plus  $x_1^h \leq \omega_1^h$  for  $h=1,2$ . The only competitive equilibrium here is individual autarky (i.e.,  $x_1^h = \omega_1^h$  for all  $h$  and  $l$ ), supported by any price  $p \geq \max\{p^1, p^2\}$ . Here  $p^h$  is the absolute value of agent  $h$ 's *MRS* at the initial endowment point, that is, the price at which  $h$  would choose zero net trades if he were not rationed.

The standard type of unrationed competitive equilibrium obtains when both agents observe both goods, and it is supported by a price  $p^*$  in the closed interval  $(p^1, p^2)$ . What if person 1 observes all net trades while person 2 observes net trades of good 1 alone? In this case, person 1 cannot be rationed and, depending on the initial endowment vector, there are two possible kinds of equilibria: an unrationed equilibrium of price  $p^*$  with agent 2 selling good 2 and buying good 1, and a rationed, individually autarkic one at price  $p^1$ .

Even in its simplest incarnation, the general equilibrium model with unobservable goods illustrates two interesting phenomena, credit rationing and the effect of observabil-

ity on price. Any of the autarkic equilibria in this section may be interpreted as credit rationing if we label goods 1 and 2 "current consumption" and "future consumption," respectively. On the other hand, if the two goods were perfect substitutes at the ratio 1:1 for consumer 1 (so that  $p^1=1$ ), then the familiar competitive equilibrium with publicly observable exchange would occur at  $p^*=1$  while the corresponding autarkic equilibrium is supported at  $p^2>1$  when trades of good 2 are not observed by anyone. Thus the unobserved commodity is more expensive than the observed one even though the two are identical in the eyes of one consumer.

The price disparity between two commodities that differ only in observability implicitly represents the cost or value of information. Unlike models of information sales, however, here information is embodied in a physical commodity and cannot be traded independently.

## III

Much recent work on efficiency wages and credit rationing stems from the observation that labor effort and future income are among the least observable net trades. This suggests that credit rationing may influence aggregate employment: credit rationed workers will typically attempt to relax tight borrowing constraints by redistributing their labor supply over time. We explore the general equilibrium implications of borrowing constraints on aggregate labor supply within the following economy. There are two time periods ( $t=1,2$ ), two perishable physical goods (bread and leisure), one asset (private loans) and three types of agents (entrepreneurs, workers of type *A*, workers of type *B*). Entrepreneurs are dummy entities that possess no endowment, make zero profit, and possess a constant returns technology converting one unit of labor services into one unit of bread. All sales revenue is distributed as wages; each period the wage rate equals one unit of bread.

Workers of type  $h=A, B$  are endowed with no bread; with a leisure vector  $(e_{1h}, e_{2h}) \geq 0$  for the two periods in which they

live, and with an additive utility function  $u^h(x_{1h}, l_{1h}) + v^h(x_{2h}, l_{2h})$  defined over their lifetime consumption of bread  $(x_{1h}, x_{2h})$  and leisure  $(l_{1h}, l_{2h})$ . Our informational assumption here is that a firm's total employment and wage bill constitute public information but the identity of its employees and the net trades of each worker remain entirely private. In symbols,  $I_{th} = \{h\}$  for all worker-commodity pairs and  $I_h = I$  for all entrepreneur-commodity pairs.

This particular information structure rules out all borrowing and lending. Workers' first-period choices are independent of second-period ones:  $u^h$  is maximized subject to  $0 \leq x_{1h} \leq l_{1h} \leq e_{1h}$  and  $v^h$  is maximized subject to  $0 \leq x_{2h} \leq l_{2h} \leq e_{2h}$ . When all trades by workers are unobservable, competitive equilibrium is a vector  $(l_{1A}^0, l_{2A}^0, l_{1B}^0, l_{2B}^0)$  subject to  $x_{th}^0 = l_{th}^0$  for  $t=1,2$  and  $h=A,B$ .

If all trades were public knowledge, then one would define familiar competitive equilibria by three demand schedules  $(L_{1h}(R), L_{2h}(R), z_h(R))$  for each worker. These describe, respectively, the worker's choice of first-period labor supply  $l_{1h}$ , second-period labor supply  $l_{2h}$ , and first-period saving  $z_h = l_{1h} - x_{1h}$ , under the standard budget constraint  $x_{1h} - l_{1h} + (1/R)(x_{2h} - l_{2h}) \leq 0$ . In this case, both wage rates equal unity and  $R$  is the (gross) interest yield on loans. The aggregate supply of loans,  $z_A + z_B$ , vanishes at the equilibrium yield  $R = R^*$ .

Suppose  $\bar{R}_h$  satisfies  $z_h(\bar{R}_h) = 0$ : it is the yield at which worker  $h$  voluntarily conforms with the Shakespearean dictum "neither a borrower nor a lender be." Assume also that  $\bar{R}_A < \bar{R}_B$ , that is, that person  $A$  would be a lender and  $B$  a borrower if the information structure permitted deferred trades. Then it is easy to see that the unrationed competitive equilibrium yield is  $R^* \in (\bar{R}_A, \bar{R}_B)$  and, furthermore,  $l_{th}^0 = L_{th}(\bar{R}_h)$  for all  $t$  and  $h$ . If the yield were  $\bar{R}_h$ , worker  $h$  would freely choose to save exactly zero and to offer the credit-rationed vector  $(l_{1h}^0, l_{2h}^0)$  of labor services, even if he faced no market-by-market constraints.

When there is a shortage of observable trades, the resulting additional budget constraints confront each worker with a shadow

yield  $\bar{R}_h$  on loans rather than with the common yield  $R^*$  that would arise if all trades were publicly observed. Since type- $A$  workers are unable to lend, the forced "reduction" in their personal yield (from  $R^*$  to  $\bar{R}_A$ ) has a substitution effect favoring period 1 consumption and leisure at the expense of period 2, and a negative income effect that tends to reduce the demand for all normal goods. For type- $B$  workers, the personal yield "rises" from  $R^*$  to  $\bar{R}_B$  causing a substitution effect that favors period 2 consumption and leisure at the expense of period 1, and a negative income effect as before. If substitution effects dominate, then borrowers will defer both consumption and leisure when they are credit rationed, while lenders will expedite consumption and leisure.

The adjustment of aggregate employment to credit rationing is a more complicated story; the conclusion depends a lot on interest and income elasticities of labor supply. For instance, when the labor supply of lenders is more elastic than that of borrowers, employment will be dominated by the reactions of lenders, falling short of the unrationed flow  $L_{1A}(R^*) + L_{1B}(R^*)$  in period 1, and exceeding  $L_{2A}(R^*) + L_{2B}(R^*)$  in period 2. Of course, this is only one of several possible outcomes.

#### IV

One property of equilibria with unobservable trades that does not depend on tastes or endowments is inefficiency. Borrowers and lenders, for instance, are unable to exploit all gains from trade and have different equilibrium marginal rates of substitution ( $\bar{R}_B$  and  $\bar{R}_A$ ) between period 1 and period 2 bread. Is this an inefficiency that a relatively uninformed government can alleviate, or can the problem only be cured by an omniscient central planner?

An omniscient central planner can always find lump sum taxes and subsidies to improve any suboptimal equilibrium. A question of greater policy relevance in economies with unobservable trades is whether government can accomplish the same objective with an *anonymous* system of transfers—that is,

using solely public information. If there is such an anonymous and beneficial intervention, we say that the original unobservable trades equilibrium is *constrained-inefficient*. For the economy of the previous section, a sensible public policy should function as a substitute for the missing credit market; it should transfer bread from type-*A* to type-*B* workers in period 1 and reverse the flow in period 2.

Consider now a transfer scheme  $T = (s, \tau)$  with the following properties: in period 1 the government sells national debt in the amount of  $2s$  units of bread, at whatever interest rate the traffic will bear. Sales proceeds are distributed as a subsidy of  $s$  per capita to all workers. In period 2, the government levies on all labor suppliers a wage tax of  $\tau < 1$  units of bread per unit of labor services, and retires all national debt. As  $T$  is completely anonymous, implementation is easy if the government observes the rate of interest and total employment in period 2. Since type-*A* workers are the ones most likely to lend to the government, this scheme transfers  $s$  units of bread per head from *A* types to *B* types in period 1 and  $\tau l_{2B}$  units of bread per capita in the opposite direction one period later. Here  $l_{2h}$  is the period 2 labor supply of worker  $h$ .

National debt provides potential lenders with a reliable "borrower" whose future income (i.e., tax revenue) is a matter of public record. Is there an intervention that makes everyone better off? The answer is yes, if the original distortion is "sufficiently large." To check, we compute how small perturbations of the transfer policy about  $T = (0, 0)$  affect the lifetime utilities  $V_A$  and  $V_B$  of each worker in equilibrium. In particular, let

$$(2) \quad V_h = u^h(\hat{x}_{1h}, \hat{l}_{1h}) + v^h(\hat{x}_{2h}, \hat{l}_{2h}).$$

Here the vector  $(\hat{x}_{1h}, \hat{x}_{2h}, \hat{l}_{1h}, \hat{l}_{2h})$  maximizes individual utility subject to the usual lifetime budget constraint  $x_{1A} - s - l_{1A} + (1/R)(x_{2A} - (1-\tau)l_{2A}) \leq 0$  for worker *A* (who trades with a reliable borrower) and subject to the added budget constraints  $x_{1B} \leq s + l_{1B}$ ,  $x_{2B} \leq (1-\tau)l_{2B}$  for worker *B*. Net-of-tax wage

rates are 1 unit of bread in period 1 and  $1-\tau$  units in period 2. The credit market clears if

$$(3a) \quad s + \hat{l}_{1A} - \hat{x}_{1A} = 2s;$$

that is, if national debt equals saving by *A* workers. The government's budget constraint is satisfied if

$$(3b) \quad 2sR = \tau(\hat{l}_{2A} + \hat{l}_{2B}).$$

This means that the government has only one independent policy instrument, say, the stock of national debt. We now compute the *total* derivative  $V_{h,s}(a)$  of worker  $h$ 's lifetime utility with respect to the transfer  $s$ , and evaluate it at  $s = a$ . By the envelope theorem and equation (3b), we obtain (in the neighborhood of  $s = 0$ )

$$(4a) \quad V_{A,s}(0) \propto 1 - 2l_{2A}^0 / (l_{2A}^0 + l_{2B}^0)$$

$$(4b) \quad V_{B,s}(0) \propto \bar{R}_B / \bar{R}_A - 2l_{2B}^0 / (l_{2A}^0 + l_{2B}^0),$$

where  $\propto$  means "is proportional to." Both of these derivatives are positive if

$$(5) \quad (2 - \theta) / \theta < l_{2A}^0 / l_{2B}^0 < 1,$$

where  $\theta = \bar{R}_B / \bar{R}_A > 1$ .

The double inequality (5) is sufficient to ensure the constrained inefficiency of equilibrium with unobservable trades. Its left-hand side means that the initial distortion, as measured by  $\theta$ , is "large enough." The right-hand side of (5) implies that the gross yield of the policy  $T$  for lenders,  $\tau l_{2B}^0 / s$ , will at least match the autarkic yield  $\bar{R}_A$  whenever sufficiently many period 2 labor suppliers are frustrated borrowers. Loosely speaking, *if the added budget constraints become tightly binding on sufficiently many people, then there are grounds for welfare-improving intervention by an uninformed government*. Well-informed governments, of course, may be able to cure relatively small distortions as well (see Bruce Greenwald and Joseph Stiglitz, 1987).

## REFERENCES

- Cooper, Russell, *Wage and Employment Patterns in Labor Contracts*, New York: Harwood Academic, 1987.
- Debreu, Gerard, *Theory of Value*, New York: Wiley & Sons, 1959.
- Greenwald, Bruce and Stiglitz, Joseph, "Welfare Economics of Economies with Imperfect Information and Incomplete Markets," mimeo., 1987.
- Hall, Robert, "Consumption," NBER Working Paper No. 2265, May 1987.
- Hart, Oliver and Holmstrom, Bengt, "The Theory of Contracts," mimeo., 1986.
- Hubbard, Glenn and Judd, Kenneth, "Liquidity Constraints, Fiscal Policy, and Consumption," *Brookings Papers on Economic Activity*, 1:1986, 1-50.

# Parallel Search and Information Gathering

By TARA VISHWANATH\*

Many microeconomic problems fall into the following category. A decision maker has a number ( $n$ ) of opportunities or projects. Each project yields an unknown reward at an uncertain time, and is characterized by an independent joint probability distribution. Once a project is selected, its reward is revealed after a random time lag when it is collected. The projects are selected sequentially in any order desired. Furthermore, at any time, a number  $m$  ( $1 \leq m < n$ ) of projects may be explored simultaneously, that is, in parallel. The decision problem is to determine the sequential strategy for choosing the projects to maximize an objective which is a function of the rewards collected.

Most of the problems in search theory, dynamic allocation problems, and many information-gathering problems fall into this class. A firm's problem of choosing technologies to develop products, job search decision of workers, price search of consumers, development of resource pools, exploration problems of mines and wells, investment decisions, and marketing strategies are but a few examples where the problem stated above is naturally applicable. Indeed, the problem is basic to many imperfect information contexts, and its analysis reveals useful insights into the nature of information acquisition and structures of markets.

The motivation for the study of parallel search and information-gathering models is obvious. In problems where information acquisition is costly and time consuming, the returns to parallel effort are higher than when undertaking a single project at a time. For example, in research and development contexts, if large improvements in information are possible at low cost in early stages of

development, then it becomes efficient to run several projects in parallel. In scheduling and dynamic allocation applications, where all projects must be undertaken, parallel operation substantially reduces the overall completion time.

The instance of the problem when only one project can be chosen at a time ( $m = 1$ ) has been studied extensively in recent economic literature. In contrast, there has been no significant attempt to study parallel selection of projects. Unfortunately, the elegant results that may be derived for single-project selection do not readily generalize to the parallel project case. Furthermore, this problem, in principle, may be formulated in a dynamic programming framework, and solved through standard techniques (such as backward induction or fixed-point methods). However, in most actual cases, this approach, besides shedding little economic insight, would be a combinatorially complex task of formidable proportions unless  $n$  and  $m$  are small. Hence, there is a need to study the effectiveness of meaningful operational rules.

The purpose of this paper is to point out that simple ordering rules are obtained for the optimal parallel selection of projects, under reasonably general and meaningful conditions. These conditions may be stated in terms of risks and stochastic orderings of the distributions associated with the projects.

The problem of single project selection ( $m = 1$ ), falls into the general class of bandit processes (see the works of J. C. Gittens, 1979, and others), the solution for which is usually characterized by a reservation rule. Each project is assigned a reservation number or an index (analogous to internal rate of return) depending only on the features of that project and independent of all other projects. At each decision instant, the project with the highest reservation number is selected. These reservation numbers, as-

\*Department of Economics, and Center for Urban Affairs and Policy Research, Northwestern University, Evanston, IL 60201. Research supported by NSF grant SES-8708325.



essed for each project in isolation, are of little relevance, in general, to the parallel projects problem which has qualitatively different features.

### I. Models and Optimal Ordering

The models described focus on search and dynamic allocation. For  $1 \leq i \leq n$ , let  $Z_i$  and  $X_i$ , respectively, denote the reward and the yield time of project  $i$ . Reward from a project is taken to be the estimated benefits net of costs involved in carrying out the project. Yield time is the time taken to collect the reward from the time the project is started. Both are nonnegative real random variables with finite expectations and having the joint distribution  $F_i(x, z) \equiv \Pr(X_i \leq x, Z_i \leq z)$ . It is assumed that  $F_i$  and  $F_j$  are independent for  $i \neq j$ .

#### A. Model I (Dynamic Allocation)

Suppose all projects must be undertaken, and at any time,  $m$  ( $1 \leq m < n$ ) projects may be carried out in parallel. The objective is to determine, at each instant, which of the projects should be in operation so as to maximize expected present value of rewards obtained from all projects. If  $t_i(\pi)$  denotes the time at which reward from project  $i$  is received following a strategy  $\pi$ , the objective is to maximize  $E[\sum \beta^{t_i(\pi)} u(Z_i)]$ , where  $\beta$  is the discount rate, and  $u(\cdot)$  denotes concave increasing utility. In this context, two cases may be distinguished depending upon whether or not there is an option to pull out of a project before its completion and switch to another. Preemptive policies admit such an option, whereas nonpreemptive policies do not. In the following only the nonpreemptive strategies are considered.

Finding the optimal parallel operation in the above model can be quite complex. However, if the reward and yield time distributions satisfy the following conditions, the optimal nonpreemptive strategy is a simple predetermined order. The three conditions are:

(C1) For each project  $k$ ,  $1 \leq k \leq n$ , given  $X_k = x$ , the conditional distribution of rewards, denoted  $F_k(z|x)$ , is either stochasti-

cally decreasing in  $x$ , or has a spread which is increasing in  $x$  (with the mean preserved).

(C2) For each project  $k$ ,  $2 \leq k \leq n$ , the conditional distribution  $F_k(z|x)$  is either stochastically smaller than, or is a mean-preserving spread of,  $F_{k-1}(z|x)$ .

(C3) For each project  $k$ ,  $1 \leq k \leq n-1$ , the yield time  $X_k$  is stochastically smaller than  $X_{k+1}$ .

The last condition may be replaced by (C3a). However, in this case, an additional restriction must be satisfied. Let  $Z_k(x)$  denote a random variable with distribution  $F_k(z|x)$ .

(C3a) For each project  $k$ ,  $1 \leq k \leq n-1$ , the yield time  $X_k$  is a mean-preserving spread of  $X_{k+1}$ . In addition  $E[u(Z_k(x))]$  is convex nonincreasing in  $x$ .

The conditions, while allowing for correlation between reward and yield times within each project, imply that, for all projects, the longer it takes to receive the reward, the less likely it will be large or that the risk involved in it is greater. In addition, the conditions order the projects in their yield times and their rewards conditional on the same yield times, either through first-order stochastic dominance or in terms of their risk. If the decision maker is risk neutral (linear utility function), the first condition is not relevant, and the second condition may be replaced by the decreasing order of the mean rewards, with project 1 having the highest.

Under the above conditions, the *optimal nonpreemptive strategy* may be stated simply: select the projects in the predetermined order  $(1, 2, \dots, n)$ . In other words, whenever an opportunity to start a project arises, from the remaining projects the one with the highest index is selected.

Given the assumptions, the expected utility of the projects, assessed individually, is decreasing in the project number, being highest for the first project and least for project  $n$ . Define

$$g_k(x) \equiv E[\beta^x u(Z_k(x))].$$

Then  $g_k(x)$  is decreasing in  $x$  due to assumption C1. Furthermore, for  $1 \leq k \leq n-1$ ,  $g_k(x) \geq g_{k+1}(x)$  for all  $x$ , due to

C2. These combined with the third assumption imply that  $E[g_k(X_k)] \geq E[g_{k+1}(X_{k+1})]$ , which in turn implies that the expected utility of the projects are ordered. Then, the optimal rule, stated otherwise is the *highest expected utility* ordering. This ordering does not depend on the specific number ( $m$ ) of parallel projects, and holds even when the number that may be operated in parallel is nondecreasing over time.

The formal demonstration of the optimality of the above rule may be found in my earlier paper (1987). A key fact is that, when the optimal ordering is followed, the expected present value of the rewards is convex decreasing in the time at which a new parallel project can be started. That is, if  $\tau_i$  denotes the time at which the  $i$ th parallel project can be started, and if  $R(\tau_1, \tau_2, \dots, \tau_m; S)$  denotes the total expected utility from carrying out the projects in set  $S$  in the optimal order, given the times  $\tau_i \leq \tau_{i+1}$ , then  $R$  is convex decreasing in each  $\tau_i$ . Thus, if project  $i$  is started earlier instead of another project  $j > i$ , then a higher reward (of project  $i$ ) is obtained sooner. In addition, the total expected utility from the other projects would also be higher, as the yield time for project  $i$  is stochastically smaller or has greater risk. The conditions stated are also tight in the sense that, if any of them are violated, the highest expected utility rule is no longer optimal.

The case of parallel projects also differs from the one project at a time case in the sense that if one considers an interchange in the order of two successive projects, the states of the other projects are unaltered in the latter, whereas the alteration occurs in the former. This observation may be used to show that the highest expected utility ordering is optimal in the case of single project selection even when new projects may become available at random times in the future. This extension does not, in general, hold when parallel projects are undertaken. Finding conditions under which simple rules are optimal, when new projects may arrive in future or when constraints on project due dates are imposed, remain open for investigation. Preemptive policies have not been

considered here. The optimality of simple rules under similar conditions may be anticipated here also.

### B. Model 2 (Parallel Search)

Suppose the projects are viewed as alternatives, each of which when explored reveals the reward that is initially uncertain, and suppose that the goal is to search for the maximum reward. While searching, several alternatives may be explored in parallel. Once the reward from a project is known, if search is continued, another one from the set of remaining projects is started. If search is terminated at any time, the maximum of the rewards received thus far is collected. The problem is to find the sequential search strategy which maximizes the expected present value of the reward at the time of stopping, net of search costs.

The projects may be viewed as substitutable technologies, the rewards for which are initially uncertain, and the search process is in essence a research and development effort from which information can be gathered, and the best technology chosen. Job search interpretations are also possible, and this will be discussed later. The one project at a time case has been studied by Martin Weitzman (1979). In this case, each project can be assigned a reservation number, and search is carried out by ordering the projects from the highest reservation number to the least. At any time, a project from the remaining set is pursued only if the current maximum of the rewards is less than the reservation numbers of all remaining projects.

In general, for the case of parallel search, the optimal rule does not have the simple structure of that of the single-project selection outlined above. However, when conditions similar to those in the previous section are imposed, a simple search order is obtained. This is the case, for example, if the distributions associated with the projects are ordered either through stochastic dominance or in terms of their risks. For a simple illustration of the ideas, suppose that there are three projects ( $n = 3$ ). Assume that the yield time of each project equals one. Let  $Z_i$

denote the random reward from project  $i$ . Let  $c$  denote the search cost (i.e., the cost per sample) of each project. Each project is sampled only once. It can be shown that, if  $Z_i$  stochastically dominates  $Z_{i+1}$  for  $i = 1, 2$ , then the search order  $(1, 2, 3)$  is optimal. To see this, consider any other order, say,  $(1, 3, 2)$ . Consider the time instant at which the first two projects have been explored. Let  $y_1$  and  $y_3$  denote the rewards observed in these two projects. At this point, if search is continued, conditional on observing  $y_2$  from project 2, the value is  $h(y_1, y_3, y_2) - c$ , where  $h$  denotes the maximum of its arguments and  $c$  is the cost that must be paid to explore the third project. Therefore, assuming no discounting, the value,  $V(1, 3, 2)$ , from searching in the order  $(1, 3, 2)$  equals

$$E[\max\{h(Z_1, Z_3), \\ h(Z_1, Z_3, Z_2) - c\}] - 2c.$$

Similarly, the value  $V(1, 2, 3)$  equals

$$E[\max\{h(Z_1, Z_2), \\ h(Z_1, Z_3, Z_2) - c\}] - 2c.$$

Now,  $\max\{x, y\}$  is increasing in both arguments, and consequently, because of the assumption,  $\max\{y, Z_2\}$  stochastically dominates  $\max\{y, Z_3\}$  for all  $y$ . It follows that  $V(1, 2, 3)$  is greater than  $V(1, 3, 2)$ . This argument which establishes the optimality of the order  $(1, 2, 3)$  can be extended for any  $n$  and  $m$ . Moreover, it can be shown that the same ordering holds even if search costs for different projects are different provided  $c_i \leq c_{i+1}$ , in other words, the costs associated with a stochastically smaller reward are greater. (Such a case arises in R&D situations involving prototype development. More about this later.) The result is not altered even if discounting during search is introduced. A similar search ordering is obtained if, instead of first-order stochastic domi-

nance, the rewards from the projects all had the same mean but different spreads. In this case, the projects with the highest spread would be pursued first. With unequal and possibly random yield times of the projects, conditions similar to those presented in the earlier section would have to be imposed to retain the optimality of the predetermined search order.

Given the assumptions outlined above, although the optimal parallel search order is simple and predetermined, the reservation numbers will not in general be the same as that for the single project selection case. They would be a function of the distributions of all projects to be pursued in parallel next. (Reservation number has the same interpretation that search is terminated if the maximum reward received to date is larger than the highest reservation number). Nevertheless, determining the optimal search strategy becomes considerably easier as the best order of project selection is known in advance.

A variant of the problem outlined above is of interest in a different context. Consider a firm which is attempting to market a product in some ( $m$ ) of several locations, the consumer demand in all of which is initially uncertain. A project, here, may be interpreted as market research in a particular location. A suitable objective for this context is the maximization of the expected value of the sum of  $m$  rewards observed thus far, net of market research costs. A predetermined search order is also optimal here, under the conditions stated for the parallel search problem outlined above.

One can inquire into the optimal number of projects that must be explored in parallel in each stage, when the number that may be searched in parallel is subject to an increasing cost. This situation corresponds to sequential search with optimal batch sampling at each stage. Several authors have studied this problem (see P. B. Morgan, 1983, and the references therein) when all projects are identical. In this context, for search with recall (i.e., when a reward received in the past can be perfectly recalled at any time), the optimal number of parallel projects

(batch size) is nonincreasing over time. My conjecture is that this property holds even in the problem of parallel search with recall and nonidentical projects described above. In search where a reward cannot be recalled later, the optimal order is typically involved (see Kenneth Burdett, 1986, who studies single project selection). Parallel search with no recall remains open for investigation.

## II. Applications

The wide scope of applicability of search, dynamic allocation, and information-gathering models is well known. In the following, I shall discuss some of the applications of the models described in the previous section.

In applications such as research and development, in the course of exploring a project, information is continually revealed about its potential benefits. When sufficient information is gathered, a decision is taken either to adopt (undertake) the project and collect the benefits or to abandon it. In the following, I shall illustrate how model 1 is applicable for the parallel exploration of such projects. For this, consider each project to be a sequential development project (SDP) as viewed in Kevin Roberts and Weitzman (1981). During the initial exploration or the development phase, an SDP passes through a number of distinct stages whose order is prescribed at the outset. For example, the first stage may be the construction of a prototype model, followed by an intensive development of a particular component, then development of another part etc. Performing each stage, which involves some cost, reveals some information about the ultimate benefits from undertaking the project. A dual decision is faced at each stage: the development process can be continued or terminated; given termination, either the project which provides the terminal benefits may be undertaken or it can be discarded. Stages of development are optional in the sense that the decision to undertake the project may be taken even before all stages are completed. Suppose that each project is assessed, that is, decision taken at each stage, so as to maximize the expected terminal benefits net of development costs. To characterize the learn-

ing of the terminal benefits of a project, let the state variable be the number of steps or stages from the end (or left to go). For simplicity, assume that the random variable  $Z_s$ , representing the terminal benefits as perceived at stage  $s$ , is normally distributed for each  $s$  with mean  $\mu_s$  and variance  $\sigma_s^2$ . As  $s$  becomes smaller, the benefits become progressively less uncertain, with the variance eventually decreasing to zero. In other words, the true benefit is learned by revising a normal prior at each stage. The time steps can be discrete or continuous. (In the latter, the learning would be a Weiner process with zero drift.) In either case, as the remaining stages decrease, the evolution of the mean  $\mu_s$  is a martingale with diminishing spread. Let  $c_s$  denote the expected cost of carrying out the remaining  $s$  stages. By postulating the relationship  $\sigma_s = \alpha c_s$ , for all  $s$ , where  $\alpha$  is a constant (i.e., higher uncertainty is associated with proportionately higher expected cost), Roberts and Weitzman (p. 1285) have shown that, if the decision maker is risk neutral, the SDP can be evaluated by using two linear symmetric stopping boundaries. Terminate development and undertake project at  $s$ , if  $\mu_s \geq as$ ; discard project if  $\mu_s \leq -as$ . The constant  $a > 0$  can be determined from the knowledge of parameter  $\alpha$ .

To see how model 1 is applicable in such a context, suppose there are several SDPs, all of which must be explored. Assume that whenever an SDP is started, it is nonpreemptively carried out to completion (i.e., it is either adopted or discarded). The yield time, or the completion time, of an SDP can be assessed before it is started based on the prior information about it. Consider an increase in the prior mean which is initially positive. Then, the distribution of  $Z_s$ , for each  $s$ , perceived at the start, improves (i.e., mean increases with shape remaining unchanged). Consequently, the probability that the SDP is terminated and adopted at any stage is increased, and the probability of discarding is decreased. This implies that the benefits conditional on the yield time improves, and that the yield time is stochastically smaller. If the SDPs have distinct and positive means, and if they all have the same level of uncertainty, then they can be ordered

from the highest mean to the lowest, with a higher mean being associated with a stochastically smaller yield time. Hence, conditions C2 and C3 are satisfied, given decision making is risk neutral (with risk neutrality, C1 is not relevant). Thus, the specified optimal ordering for parallel exploration is applicable.

Search models have been used to characterize the behavior of workers in labor supply and unemployment studies (see D. T. Mortensen, 1984). In the literature on systematic job search, an unemployed worker must order his employment prospects to maximize his expected income. Employment prospects that a worker explores may be interpreted as projects. Steven Salop (1973) has analyzed the case where a worker explores only one prospect per period. Model 2 is directly applicable to describe the behavior of a worker engaged in parallel exploration of the prospects. Here, it can be shown that the reservation wage is negatively dependent on the length of unemployment, an empirically observed fact.

The search model also arises in the context of a firm or a central planner faced with the choice of alternative substitutable technologies to develop a product. Each project may be interpreted as the development and research required to assess the benefits of a technology. Early work on parallel choice may be found in the studies of R. R. Nelson (1961), and Jacob Marschak et al. (1967). They discuss the importance and the associated difficulties. The results outlined for model 2 provides simple guidelines for this problem of parallel search.

The above discussions focus on the decision making of a single agent. In the following, some implications of the results in the context of a game involving several agents is discussed. Specifically, I shall consider  $N$  firms competing to develop a new product, and show that the Nash equilibrium strategy is a simple rule. Suppose that all firms have access to the same set of  $L$  substitutable technologies. A project for a firm is the exploration of a technology which reveals (say, in one period) its unit production cost which is initially uncertain. Let  $Z_{ik}$  represent the random unit production cost for

firm  $i$ , from technology  $k$ . Suppose that these firm and technology-specific costs satisfy the monotonicity assumption: for all  $i$ , the cost  $Z_{ik}$  is stochastically smaller than  $Z_{ij}$ , if  $k < j$ . This ordering common to all firms may be justified if they have the same technology specific background information. Further, assume that the search (exploration) cost for firm  $i$  is such that  $c_{ik} < c_{ij}$ , for  $k < j$ . (This is appropriate when exploration involves construction of prototypes.) Each firm engages in parallel search for the minimum-cost technology. The ultimate reward for firm  $i$  is proportional to the Cournot-Nash equilibrium production level  $q_i(c_1, c_2, \dots, c_N)$ , which depends on the cost chosen by all firms. Because,  $q_i$  decreases in  $c_i$ , given any search strategy  $\alpha_i$  of the opponents of firm  $i$ , the reward  $q_i(Z_{ik}; \alpha_i)$  from project  $k$  is stochastically larger than  $q_i(Z_{ij}; \alpha_i)$ , for  $k < j$ . From the results of model 2, it follows that the Nash equilibrium parallel search order is  $(1, 2, \dots, L)$  for all firms. Jennifer Reinganum (1983) has derived the optimality of this search order for the case when each firm chooses a single project at a time. Following the analysis there, the existence of the Nash equilibrium reservation numbers (stopping criterion) can also be shown for parallel selection. It must be emphasized that the monotonicity assumption above is restrictive, and may not be satisfied if firms have differential initial information about the technologies. For this case, the existence of Nash equilibrium and whether the strategies have the same simple structure as that of single-agent problems remain to be investigated. Finally, the results for parallel activity derived here may also prove useful in evaluating the efficiency of competitive situations in R&D environments.

### III. Conclusions

Parallel search and information-gathering problems, in general, are computationally complex, and there has been relatively little research effort in this direction. This paper shows that under some meaningful conditions, simple ordering rules are obtained for the optimal strategy. Although the condi-

tions imposed are somewhat restrictive, this study may be a starting point for finding less stringent conditions which yield similar simple rules of interest in applications.

#### REFERENCES

- Burdett, K., "Systematic Search: The Case of No Recall," mimeo., Cornell University, 1986.
- Gittens, J. C., "Bandit Processes and Dynamic Allocation Indices," *Journal of Royal Statistical Society*, 1979, 41, 148-77.
- Marshall, J., Gilenon, T. O. and Summers, R., *Strategy for R&D*, New York: Springer-Verlag, 1967.
- Morgan, P. B., "Search and Optimal Sample Sizes," *Review of Economic Studies*, October 1983, 163, 659-76.
- Mortensen, D. T., "Job Search and Labor Market Analysis," in R. Layard and O. Ashenfelter, eds., *Handbook of Labor Economics*, Amsterdam: North-Holland, 1984.
- Nelson, R. R., "Uncertainty, Learning, and the Economics of Parallel R and D Efforts," *Review of Economics and Statistics*, November 1961, 43, 351-64.
- Roberts, K. and Weitzman, M., "Funding criteria for Research, Development, and Exploration Projects," *Econometrica*, September 1981, 49, 1261-88.
- Reinganum, J., "Nash Equilibrium Search for the Best Alternative," *Journal of Economic Theory*, June 1983, 30, 139-52.
- Rothschild, M. and Stiglitz, J., "Increasing Risk: A Definition," *Journal of Economic Theory*, September 1970, 9, 185-202.
- Salop, S., "Systematic Job Search and Unemployment," *Review of Economic Studies*, July 1973, 40, 191-201.
- Weitzman, M., "Optimal Search for the Best Alternative," *Econometrica*, May 1979, 47, 637-54.
- Vishwanath, T., "Optimal Ordering for Parallel Search," mimeo., Northwestern University, 1987.

## CHALLENGING SOME CONVENTIONAL WISDOMS ABOUT THE LABOR MARKET<sup>†</sup>

### The Un-Natural Rate of Unemployment: An Econometric Critique of the NAIRU Hypothesis

By DAVID M. GORDON\*

Few ideas have taken such firm hold in neoclassical economics as the concept of a "natural rate of unemployment" (see my 1987 paper). The transformation of the prevailing wisdom is compactly illustrated by the evolution of the profession's exemplary elementary textbook, Paul Samuelson's *Economics*. In his 8th edition, Samuelson defined "full employment as a condition where 96 1/2 percent of the labor force are employed, rather than where only 94 or 95 percent are employed" (1970, p. 801). By the 12th edition in 1985, reflecting the new conventional wisdom, Samuelson (with William D. Nordhaus) reports: "Modern mainstream macro says that there is a natural rate of unemployment—today around 6 percent—below which the economy cannot go without running the straits of inflation" (p. 766).

Reporting on work-in-progress, I argue that this "modern mainstream macro" view is unwarranted. (See my 1988 paper for a more complete presentation.) I concentrate here on only two strands of the argument. I first present econometric evidence that, *even on its own terms*, the natural rate hypothesis does not provide a convincing explanation of the deteriorating tradeoff between inflation

and unemployment in the United States from the 1960's to the 1980's. I then suggest that the basic models underlying that hypothesis, upon which the hoary visions of accelerating inflation are grounded, themselves provide substantially incomplete representations of the dynamics of wage and price determination and consequently suffer from under-specification bias. Once one takes account of structured mediations of capital-labor conflict and dynamic interactions between economic growth and pricing strategy, prevailing conclusions of a vertical tradeoff between unemployment and inflation begin to dissolve.

#### I. The Econometrics of a Rising NAIRU: Where's the Beef?

Some argue that the natural rate of unemployment has increased simply because the demographic composition of the labor force has changed. This argument is incomplete and essentially tautological, since by itself it embodies no stochastic tests of the link between these demographic changes and changing dynamics of wage and price determination.

I have therefore concentrated on direct stochastic applications of models of the natural rate hypothesis, focusing on notions of an "equilibrium" or "non-accelerating-inflation" rate of unemployment (NAIRU). (See G. E. Johnson and P. R. G. Layard, 1986, for a general review.)

This method is most compactly summarized for a reduced-form inflation equation in which the dependent variable in a nominal wage-change equation is substituted as an independent variable into a price-

<sup>†</sup>*Discussants:* Orley Ashenfelter, Princeton University; Marvin Kosters, American Enterprise Institute; Robert Topel, University of Chicago.

\*Department of Economics, Graduate Faculty, New School for Social Research, 66 W. 12th St., New York, NY 10011. I am grateful to Robert J. Gordon for sharing his data and to members of the Center for Democratic Alternatives Macro Working Group for helpful discussion. This work was partly supported by a research grant from the National Science Foundation.

change equation, resulting in a reduced-form inflation equation in which current inflation is a function of expected inflation, a demand-pressure variable, and a variety of other "supply-side" influences on wage-price dynamics. Relying on Robert J. Gordon's elaboration and notation (1985, pp. 267-71), this reduced-form equation for price change over time can be written as

$$(1) \quad p = p^e + m(v) + (1/a + b)[hgX + z],$$

where time subscripts are suppressed, lower-case letters designate time derivatives,  $P$  is the product price;  $P^e$  is the expected product price;  $M$  is the markup expressed as a function of  $V$ , measuring conditions of excess demand in the commodity market;  $X$  is a measure of excess labor-market demand (for example, of "excess" unemployment);  $Z$  is a vector of other supply-side influences;  $a$  and  $b$  are, respectively, the real-wage elasticities of labor demand and supply;  $g$  is the coefficient of dynamic labor-market adjustment to the gap between demand and supply; and  $h$  is the proportion of domestic inputs in total final product. As written, the coefficient of 1.0 on  $p^e$  indicates that equation (1) satisfies the conditions of perfect adjustment of current to expected rates of price change.

The definition of the NAIRU follows directly from (1). At a constant level of excess demand ( $v = 0$ ), inflation accelerates ( $p > p^e$ ) whenever  $X$  is positive and decelerates whenever  $X$  is negative. If we (i) use the rate of unemployment ( $U$ ) as a measure of labor-market demand pressure and designate  $U^*$  as the NAIRU; (ii) allow a constant ( $c_1$ ) to appear in the equation; (iii) denote  $[-hg/(a + b)] = c_2$ ; and (iv) impose the restriction that  $p = p^e$ , we can then use (1) to solve for  $U^*$ :

$$(2) \quad U^* = [c_1(a + b) + z] / -hg;$$

$$\text{or} \quad U^* = c_1/c_2 \quad \text{when } z = 0.$$

If  $z$  were equal to zero, we get what R. J. Gordon calls the "no-shock" natural rate (hereinafter abbreviated as NAINORU), or  $U^{*N}$ . If supply "shocks" are manifest

(and, therefore, accommodated), we get what I would call the non-accelerating-inflation, shock-accommodating rate of inflation (NAISARU), or  $U^{*S}$ .

Gordon has provided what appear to be the most robust estimates of these relationships for the United States (see 1985 and references to other work therein). According to his specification and estimation, the NAINORU increased from roughly 4.5 percent in the 1950's and 1960's to roughly 6.0 percent in the 1980's. In order to set aside (as much as possible) issues of variable definition and data commensurability, I have based all of the discussion in this section on Gordon's data and specifications.

It is worth stressing at the outset that a rising natural rate of unemployment does *not* follow automatically from this approach. Most concretely, if we use the measured aggregate unemployment rate as the direct measure of labor-market demand pressure in (1) and assume that there are no supply shocks, then the NAINORU is *constant* by definition, with  $U^* = c_1/c_2$ , and can therefore not have increased during the period of estimation.

Working within this approach, however, one might still find at least three alternative kinds of econometric evidence for a rising NAIRU.

#### A. A Weighted Unemployment Rate

Instead of the actual rate of unemployment as a labor-market demand-pressure variable, some use an unemployment rate weighted for changing demographic composition ( $U^W$ ) on the grounds that the labor market works more or less smoothly depending on the (weighted) "quality" of the labor force. If  $U^W$  is substituted for  $U$  in (1) and (2), then an estimate for  $U^{*W}$  can be derived from (2), still necessarily *constant* over time, while a variable estimate for  $U^*$  is then derived by an algebraic transformation based on the variable difference between the actual and weighted unemployment rate:  $U^* = U^{*W} + (U^W - U)$ . This is the basis for Gordon's estimate of a rising NAINORU since the mid-1950's, since it does turn out that ( $U^W - U$ ) itself increased.



There are two principal problems with this method. First, the weighted-unemployment rate, following George Perry's lead (1970, especially pp. 439–40), is itself constructed by using relative wages as weights for aggregating individual age-sex groups' share of the total labor force. This presupposes that (a) relative wage rates accurately and monotonically reflect differences in relative productivities; (b) those relative weights are stable over time; and (c) relative labor market demand does not adjust sufficiently smoothly to (or, indeed, even partly cause) shifting age-sex labor-supply composition so that labor-market friction can be taken to be a direct function of the demographic composition of the labor force. All of these presumptions seem sufficiently problematic that one should avoid their purely axiomatic presupposition and therefore not lean exclusively or too heavily on this particular variable construction.

Second, this tactic implies that the weighted-unemployment rate is a better measure of labor-market slackness than the actual unemployment rate—presumably because the latter is more fraught with errors-in-variable than the former. But the econometric evidence supporting this estimating tactic is weak at best. In the reduced-form specification of (1), the explanatory power of an equation using  $U^w$  is empirically indistinguishable from one using  $U$ . (In a test for differences in the residual sum of squares of the two equations,  $F < 1.00$ .) Since  $U^w$  otherwise builds in behavioral presuppositions which should be tested, not merely presupposed, it would seem unwarranted to base an estimate of the NAIRU on an equation making use of  $U^w$  rather than  $U$ . But without further specification, it is *only* the former, not the latter, which permits an estimate of a *rising* NAIRU.

### B. Changing Slope Coefficients?

One direct test of expectations about a rising natural rate builds from the idea of increasing wage rigidity. Translated econometrically, this would suggest piecewise regression tests for changes in the slope of  $X$  in (1), potentially indicating that a unit re-

duction in the level of unemployment (or another demand-pressure variable) would result in a relatively greater increase in inflation as the postwar period progressed.

There is no evidence of such an effect. Using both the reduced-form equation and the separate (structural) inflation and nominal-wage-change equations, I have tested for piecewise changes in the regression coefficients on the unemployment rate in the successive business cycles after 1973. Not one of the unrestricted lag coefficients for either cycle in any of the three equations was statistically significant.

### C. Upward Shifts in the NAIRU Curve?

If, on the basis of these first several tests, one might reasonably conclude that the NAIRU has been constant over time, it might nonetheless be the case that there had been an "outward shift" in the relationship between inflation and unemployment over time for which the natural rate hypothesis might plausibly provide a proximate explanation—even if not captured by shifts in the slope coefficients on  $U$  itself. At the first level of approximation, this would require a statistically significant upward shift in the intercept of (1) since the 1960's.

There is, perforce, evidence of such a shift. Using separate dummy variables for Gordon's periodization (1973–76, 1977–80, 1981–84) or for a business cycle periodization (1973–79, 1979–84), I find evidence of a statistically significant two-stage upward shift in the intercepts of the reduced-form equation.

The problem with this result is that the intercept shifts seem to reflect quite different dynamics than are conventionally advanced by the NAIRU analysis. Shifting from the reduced-form to the separate inflation and wage-change equations, I find that the entire upward shift is strictly confined to the inflation equation. If there has been an upward shift in the NAIRU curve, by these indicators, it would appear to reflect unspecified changes in pricing behavior (in the markup over unit costs), not a diminishing flexibility of wage response to relative labor-market pressure.

## II. The Dynamics of Perfect Adjustment: Where's the Vertical Line?

All the foregoing discussion reflects solely on the evidence for a rising natural rate of unemployment over a specific period of history. It does not yet address the central microfoundations of the idea of a natural rate. Even if the preceding econometric evidence suggests that we have been operating *above* a constant NAIRU over the past 15 years, with inflationary pressures from other sources, this in no way compromises the prevailing expectation that a reduction in the unemployment rate below the NAIRU, whatever its current level, would immediately trigger accelerating inflation.

I therefore arrive at the theoretical punchline of the natural rate proposition: it is based on the notion that there is full adjustment of prices to wages (and vice versa) with no long-run tradeoff between inflation and unemployment.

One of the strongest strands of empirical evidence for a vertical long-run tradeoff between inflation and unemployment, indeed, was that earlier Phillips curve equations had been underspecified, inattentive to a variety of additional supply-side determinants of wage-price adjustment; the early Phillips curve coefficients, normally falling significantly below 1.00, apparently suffered almost fatally from underspecification bias. The econometrics could hardly have been better scripted if Milton Friedman or Edmund S. Phelps had generated the data themselves.

But are more recent models of wage-price adjustment any less vulnerable to such complaints?

Upon reflection, it would be surprising if it were *not* possible to improve upon prevailing accounts. In the NAIRU literature, most empirical treatments of price adjustment build from the sparest of markup models. And almost all analyses of nominal wage change begin and end with only the faintest nod toward structured bargaining models or the dynamics of capital-labor conflict. But the issue for my current purposes is not so much the simplicity of prevailing specifications as the robustness of their empiri-

cal estimates of the unemployment-inflation tradeoff when confronted with more textured analyses.

Reporting on work-in-progress, I conclude that recent estimates supporting no long-run tradeoff between unemployment and inflation suffer from the same disease as their earlier Keynesian predecessors: they are contaminated with underspecification bias. Constrained by space limitations I summarize here only the most skeletal outlines of an alternative account of wage and price determination—reviewing just enough of its results to suggest the shakiness of the prevailing NAIRU perspective. (See my work-in-progress both for more detail of argument and for references to the literature on which this analysis builds.) Compared with more fully articulated models of wage and price dynamics, it appears the vertical Phillips curve begins to look like the leaning tower.

### A. Wage Determination

Prevailing mainstream models of wage determination seem incomplete along theoretical, institutional, and historical dimensions. I postulate an expanded model of wage determination which seeks formally to integrate four main additional analytic concerns: 1) At the most general level, employers set a target nominal wage in order to achieve the optimal level of labor intensity, while workers seek to achieve a target rate of real wage increase. 2) The relative influence of these respective objectives is mediated by the "reserve army effect," by the relative disciplinary effectiveness of external labor-market conditions. 3) Structured wage bargaining, primarily in enterprises featuring union representation, is additionally conditioned by the specific dynamics of labor-management expectations—in the United States, for example, primarily by the system of "productivity bargaining." 4) The trajectory of wage bargaining in the United States was critically affected after the early 1970's by increasingly intensive "employers' offensives," themselves spurred by the profit squeeze of the late 1960's.

These four elements can be unified into a single operational model of the determina-

tion of nominal wage change. Estimated for the United States over the postwar period, the model appears both to receive empirical confirmation for each of its constituent elements and significantly to improve the explanatory power of mainstream models of wage determination. The implications of those empirical results are summarized below.

### B. Price Determination

I have also sought more fully to elaborate models of price determination. Beginning as in many prevailing mainstream accounts from a foundation of markup pricing, I have attempted further to incorporate two additional dimensions of determination. 1) Traditional markup models pay insufficient attention to the richness of both short-term and longer-term influences on variations in competitive pressure and therefore on the level of the markup over unit costs. 2) Traditional markup models have paid too little attention to target pricing models and, as a result, to interactions between price decisions and expected market growth.

It is possible to attend to these concerns within the general framework of markup pricing models and thus to compare their relative importance with somewhat more standard approaches. I have also estimated such an extended model for the postwar United States; it also appears to receive strong confirmation in the data.

### C. The Long-Run Tradeoff

I shall review here only one aspect of the empirical results of these models: their implications for estimates of the long-run tradeoff between inflation and unemployment. As noted above, the natural rate hinges critically on the coefficient on expected inflation-wage change in the relevant wage-change, inflation, or reduced-form equations. What happens to that coefficient in more fully specified models?

Table 1 presents a compact report. Two equations are summarized: an inflation equation, with expected nominal wage change as an independent variable; and a nominal

TABLE 1—THE EFFECTS OF ALTERNATIVE SPECIFICATION OF INFLATION AND NOMINAL WAGE-CHANGE EQUATIONS: RESULTS FOR THE UNITED STATES, 1954:II–1986:III

Stage of Estimation	Inflation Equation $\hat{\beta}_w$	Wage Equation $\hat{\beta}_p$
1) Original Model	1.021	1.135
<i>Technical Modifications</i>		
2) Adjust control vars.	1.053	1.087
3) Adjust behav. vars.	1.012	0.777
<i>Substantive Modifications</i>		
4) Long-run competition	0.848	
5) Expected demand growth	0.683	
6) Short-run bottlenecks	0.415	
7) Firm target wages		0.621
8) Worker target wages		0.543
9) Union bargaining		0.358
10) Surplus labor effects		0.330
11) Management offensive		0.553

wage-change equation, with expected inflation as a right-hand side variable. For ease of interpretation, the equations begin with the precise specifications deployed by Gordon in his work, now estimated through 1986, and then report on the effects of an iterative sequence of reestimated equations, with each iteration additively and cumulatively modifying the previous specification.<sup>1</sup> The first stage reports on a series of technical modifications necessary to translate from the benchmark models to specifications warranted by my own alternatives. The second stage reports on the effects of each of the substantive behavioral additions outlined above. The rows report on each sequential phase of model modification, while the single

<sup>1</sup>The specifications reported in the table as the "original model" differ from those reported in Gordon (1985) in a few minor respects: they report on estimations for separate wage-change and inflation equations rather than the single reduced form; they are estimated through 1986:III on Gordon's data (incorporating NIPA revisions) instead of through 1984:IV; they use polynomial distributed lags on the expected inflation and wage-change independent variables instead of step rectangular lag distributions; and they incorporate the actual unemployment rate as the demand-pressure variable in both equations. None of these changes significantly alters the results for the original model.

column for each equation reports the sum of the estimated coefficients on the terms for expected nominal wage change and inflation, respectively.

Minor changes are involved in the phase of "technical modifications." The adjustment of independent variables (row 2) involves some cleaning out of insignificant variables from the original specifications and some transformation of a few other supply-side variables. The adjustment for "behavioral variables" (row 3) consists primarily of changes in the wage and productivity variables from detrended indices for all employees to measures expressed for production workers which are not detrended. With the exception of the "behavioral variable" adjustment in the wage-change equation, these technical modifications do not have much impact on the performance of the equations and in any case account for none of the change in the expected-wage coefficients in the inflation equation and only part of the corresponding change in the nominal-wage equation.

Once one moves to the "substantive modifications" (rows 4-6 and 7-10), one finds that the estimated coefficients on expected wage change-inflation drop markedly. The models' explanatory power also improves: the (unreported) adjusted  $R^2$  increases from .822 for the original inflation equation (row 1) to .899 in row 6 and from .846 for the original wage-change equation (row 1) to .911 in row 10.

Why do these estimated adjustment coefficients now fall so far below 1.00? These models and results are very preliminary and, at this stage, little more than exemplary. But I nonetheless reach some provisional conclusions about their internal dynamics. In the inflation equation, wages are only partly passed on for two main reasons: when demand growth is rapid, firms limit price increases in order to hold on to or expand their market shares; and when demand growth is slow, supply-side bottlenecks develop which either reduce competitive pressure or increase effective unit costs. In the wage-change equation, two factors seem to play the largest role in moderating the effect

of price changes on wage changes: 1) Firms' target wages are defined solely with respect to labor extraction strategies and not with respect to expected product prices; while workers' target real wage demands are likely to moderate if recent real wage growth has been especially rapid; and 2) wages in the union sector both appear to exhibit some inertia and are modulated by fluctuations in the real wage share through "productivity bargaining."

### III. Conclusion

Many if not most mainstream economists have believed for some time that rising levels of unemployment in the United States were warranted by and largely reflected a "rising natural rate of unemployment." The pervasiveness of these beliefs has reflected in part the widespread fascination with "new classical" presumptions about instantaneous adjustment in product and labor markets. But their widespread acceptance has also reflected, to some large degree, a broadening acceptance of econometric studies alleging to support the NAIRU hypothesis.

The research summarized in this paper, although still provisional, would appear to cast considerable doubt upon those econometric foundations. And without such econometric support, the belief that we cannot reduce unemployment in the United States below 6 percent (without "running the straits of inflation") would appear to reflect little more than revealed and entirely contestable theoretical preference, not hardened and unyielding empirical fact.

### REFERENCES

- Gordon, David M., "Six-Percent Unemployment Ain't Natural: Demystifying the Idea of a Rising 'Natural Rate of Unemployment'," *Social Research*, Summer 1987, 54, 223-46.
- \_\_\_\_\_, "'But the NAIRU Has No Clothes': Institutions, Expectations, and the Inflation/Unemployment Trade-off," paper in progress, New School for Social Research,

1983.

Gordon, Robert J., "Understanding Inflation in the 1980s," *Brookings Papers on Economic Activity*, 1:1985, 263-99.

Johnson, G. E. and Layard, P. R. G., "The Natural Rate of Unemployment: Explanation and Policy," in O. Ashenfelter and R. Layard, eds., *Handbook of Labor Eco-*

*nomics, Vol. II*, New York: North-Holland, 1986, 921-99.

Perry, George L., "Changing Labor Markets and Inflation," *Brookings Papers on Economic Activity*, 3:1970, 411-41.

Samuelson, Paul and Nordhaus, William D., *Economics*, New York: McGraw-Hill, 8th ed., 1970; 12th ed., 1985.

# The Growth of Low-Wage Employment: 1963–86

By BARRY BLUESTONE AND BENNETT HARRISON\*

No one denies that, for well over a decade, the United States has been in the enviable position of producing more new jobs than virtually all of the other industrialized nations combined. Europe had essentially zero job growth between 1973 and 1986, while the United States created nearly 26 million net new jobs during the same period. As a result, America has been rightfully called "The Great Jobs Machine."

What is in question, however, is the *quality* of the employment generated during this period. The evidence presented here suggests that the proportion of low-wage employment has grown substantially, particularly since the late 1970's, and that at least among year-round full-time workers, there has been a tendency toward wage polarization. In this decade, there has been a significant increase in low-wage employment despite the steady recovery of the economy since 1982, and the growth in employment has increasingly been concentrated in the tails of the distribution.

## I. The Great Wage Debate

The controversy over "the low-wage question" began soon after the U.S. Joint Economic Committee released a report in December 1986 claiming that the share of net new employment paying low wages had increased significantly after 1979 (see our 1986 paper). The initial report indicated that between 1979 and 1984 the share of net new employment paying wages below one-half of the 1973 real median annual wage was nearly three-fifths, in contrast to the 1963–79 period when the proportion was less than one out of five. These results were met with immediate skepticism. For example, Janet Norwood, Commissioner of the U.S. Bureau

of Labor Statistics, noted that lack of progress toward reducing low-wage employment "reflects the impact of the 1981 to 1982 recession rather than a general inability to generate good jobs" (1987).

In a more recent treatment, Marvin Kosters and Murray Ross (1987) subjected the same *Current Population Survey (CPS)* data as were used in the JEC study to a series of sensitivity tests. Using data for every year between 1967 and 1985, slightly different wage cutoffs for defining "low" and "high" wages, substituting the Bureau of Labor Statistics' experimental CPI-X deflator for the regularly published all-item CPI, and using an estimated smoothed median as a replacement for the raw median, Kosters and Ross concluded that "compared with the JEC study, our analysis shows no rise in the share of new jobs with low annual earnings, and a significant rise instead of a decline in the share with high earnings" (p. 22). Essentially, Kosters and Ross claim, the low-wage thesis appears to be based on pure statistical artifact.

In reviewing the critique of the low-wage thesis, it is evident that there are three separate criticisms, two of which are interrelated. First, by restricting the end point of the analysis to 1984, and reporting only on particular years (1963, 1973, 1979, and 1984), much of what passed for secular trend in the JEC paper might merely have been the consequence of business cycle effects. Second, by including in the original study *all* workers—part-time and part-year workers as well as those employed year-round full-time (*YRFT*)—any secular trend in low-wage employment was potentially contaminated by a sharp cyclical component in annual earnings emanating from variations in annual hours worked. Indeed, analysis of the original JEC results indicates that more than 90 percent of the net additions to the low-wage labor force between 1979 and 1984 were part-time or part-year workers. Hence, over the business cycle, the level and distri-

\*University of Massachusetts-Boston, Boston, MA 02125, and Massachusetts Institute of Technology, Cambridge, MA 02139, respectively.

bution of annual earnings will reflect layoffs and short workweeks as well as any changes in compensation per unit of time employed. Finally, in using the BLS official CPI rather than an adjusted CPI, the original JEC analysis overstated the decline in real wages after 1979 because of distortions in the inflation index owing to BLS methods of treating owner-occupied housing costs prior to 1983 (Kosters and Ross).

## II. A Reappraisal of the Low-Wage Thesis

To deal with the first two and most important criticisms, we have reconstructed the original JEC analysis in the following manner. Based on newly available data, the *CPS* sample has been extended through 1986, adding two more economic expansion years to the time-series. Moreover, the analysis has been restricted to year-round full-time workers,<sup>1</sup> and the low-, middle-, and high-wage shares of employment have been calculated for every year between 1963 and 1986. The resulting time-series of low-wage shares is then subjected to a standard decycling procedure using a variety of business cycle measures to test for the robustness of results. This is followed by a regression analysis designed to test hypotheses about the significance of several widely held explanations for the growth of low-wage employment.

The procedure is as follows. 1) The median annual earnings for all year-round full-time workers age 16 or over was calculated for 1973.<sup>2</sup> 2) Following the JEC methodology, a low-wage cutoff was arbitrarily assigned at 50 percent of this median. Similarly, we assigned a high-wage cutoff at 200 percent of the median. 3) Low and high real wage cutoffs for 1963 to 1986 were then calculated by adjusting the 1973 cutoffs to the all item

CPI for each year.<sup>3</sup> 4) The proportion of all year-round full-time workers falling into each earnings stratum was tallied for each year. 5) Finally, to remove any business cycle component, the resulting time-series was decycled according to six different variables: real *GNP*,  $\ln GNP$ , unemployment rate,  $\ln$ unemployment rate, the capacity utilization rate, and  $\ln$ capacity utilization.<sup>4</sup>

## III. The Results

Before decycling, the raw low-wage share shows a sharp decline from 21.4 percent of the total *YRFT* workforce in 1963 to 12.5 percent in 1970; it fluctuates in a narrow band between 12.5 and 13.9 percent between 1970 and 1979; and then expands rapidly to 17.2 percent by 1986 (see Figure 1). Put simply, the low-wage share has taken a great U turn, with over half of the improvement in the low-wage share between 1963 and 1970 reversed by 1986. No matter which variable is used to decycle the trend, the U turn remains strongly in evidence. Of the six decycling variables examined here, only  $\ln GNP$  was statistically significant ( $t = 3.21$ ), and it explained only 29 percent of the variance in the low-wage earnings share.<sup>5</sup> Indeed, the rise in the decycled low-wage share since 1979 (regardless of the decycling variable chosen) is greater than the trend in the raw data, suggesting that the low-wage share has been increasing despite of and independent of four years of economic expansion and falling unemployment.

The trends in the size of the middle- and high-wage strata are equally striking, the middle stratum share falling from nearly 81 percent in 1970 to only 74 percent in 1986. In contrast, the high stratum share of *YRFT* employment peaked in 1973 at 9.1 percent,

<sup>1</sup>Year-round full-time workers comprise roughly 55 to 60 percent of the total employed labor force. As expected, the proportion varies cyclically, ranging from 53 percent in the recession year 1975 to 59 percent in 1985.

<sup>2</sup>The year 1973 was chosen as it represents the postwar year in which real average weekly earnings peaked. Thus, we are comparing wage shares against the year in which real wages were a maximum.

<sup>3</sup>This produces real wage cutoffs (in 1986 dollars) equivalent to: low-wage < \$11,104; Middle-wage \$11,104–\$44,412; High-wage \$44,413 +.

<sup>4</sup>This is accomplished in a two-step regression procedure of the following form: 1) Decycling variable<sub>*t*</sub> =  $a_0 + a_1 \text{Time} + e_{1t}$ , and 2) low-wage share<sub>*t*</sub> =  $b_0 + b_1 e_{1t} + e_{2t}$ .

<sup>5</sup>The remaining five decycling variables explained a negligible share of variance in the low-wage earnings trend and in no case had a *t*-value in excess of 0.8.

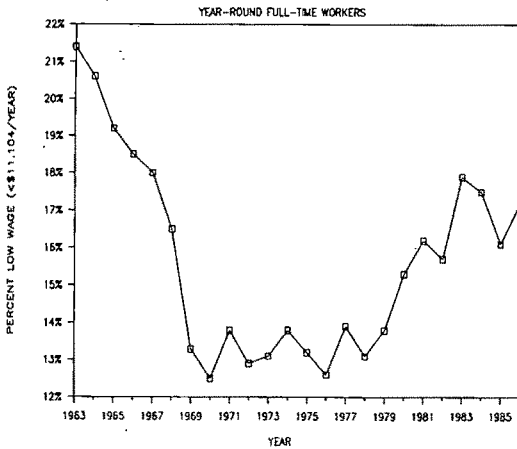


FIGURE 1. LOW-WAGE EMPLOYMENT SHARE

fell to as low as 6.5 percent during the recession year 1981, and then recovered to 8.8 percent by 1986. Taken together, the three trajectories indicate an acute increase in wage polarization since the late 1970's.

There is also evidence that using alternative deflators and medians does not substantially alter these results. Kosters and Ross focus their paper on trends among all workers. But, when they do turn their attention to *YRFT* employment, they produce results that corroborate the U turn in low-wage work, despite the fact that they employ the CPI-X deflator and use the estimated smoothed median for calculating wage cut-offs. According to their own calculations, the *YRFT* low-wage share declines from 17.2 percent in 1967 to 10.1 percent in 1979 and then rises nearly steadily to 13.8 percent by 1985 (see their Table 14, p. 39). In fact, the Kosters-Ross results actually indicate a much steeper proportional rise in the low-wage share between 1979 and 1985 than do our own calculations.

#### IV. The Demographic and Structural Parameters of Low-Wage Growth

Further analysis of the *CPS* survey data reveals that the U turn in the low-wage employment trend and the drift toward polarization are found among most demographic groups, across most regions, and

within both manufacturing and service industries. The few exceptions are instructive.

The pattern of a shrinking middle found for the *YRFT* workforce as a whole is mirrored among men and women, and generally among whites and nonwhites. What differences do exist are in degree rather than direction. The pattern is most exaggerated among men, among younger workers (age 20-34), and among those who graduated from high school but did not go to college. It is attenuated among women and those with some college. It is weakest among workers aged 35 to 54.

Regionally, the most exaggerated pattern is found in the "deindustrializing" Midwest where unemployment rates have remained high in spite of the recent economic recovery and where high-wage manufacturing employment has declined absolutely. The Northeast and particularly the New England region have experienced the slowest growth in low-wage employment and the greatest growth at the high end of the distribution. As far as industrial sectors are concerned, we do find that manufacturing has undergone the same type of U turn as the service economy (including wholesale and retail trade).

These results suggest that all groups and all sectors of the economy are subject to the same qualitative phenomena of increasing low-wage employment and polarization. But the magnitudes do vary among the various groups. In particular, at this univariate level of analysis, there does appear to be a cohort effect, with younger workers in the baby-boom generation suffering the largest increase in low-wage employment. If this reflects simply a lack of work experience, then one would expect this younger group to "catch up" with the generation ahead of them as they mature in the market (Robert Lawrence, 1984). If, on the other hand, this represents the "crowding" of a generation of workers into a labor-market structure offering lower wage opportunities, then as Frank Levy and Richard Michel (1986) suspect, this large generation of workers may be condemned permanently to lower wages and significantly attenuated age-earnings profiles.

The fact that virtually all of the expected job growth in the future will be in the service



economy provides at least some *prima facie* evidence that Levy and Michel's forecast may prove more accurate. That the low-wage share of service sector employment is nearly 22 percent compared with 12.4 percent in manufacturing and 17.2 percent among all *YRFT* workers suggests that unless there is a dramatic increase in wage levels within this sector (reflecting increases in productivity and in the bargaining power of service employees), the new labor force that enters the market will face lower wages than the older generation with seniority privileges in existing jobs.

The superior performance of the New England economy and the rapid deterioration of wages in the midwest suggest two related factors at work. One is the potentially powerful impact of deindustrialization in the heavy manufacturing sector; the other is the critical importance of sustained full employment. The loss of manufacturing jobs affects the Midwest more than any other single region of the nation, while low unemployment in New England has forced virtually all employers there—and particularly those in the service and retail trade economy—to enter a "bidding war" to attract employees.

#### V. Testing the Determinants of Low-Wage Proliferation

To ascertain the multivariate impact of these various factors on the low wage trend, a reduced-form double-log<sub>e</sub> time-series regression was performed over the period 1963–86. The GNP-decycled *YRFT* worker low-wage share was regressed on three variables. These included 1) a standard measure of annual productivity—output per person hour in all business establishments ( $\ln Q_{perH}$ ); 2) the proportion of the employed labor force aged 25 to 34 ( $\ln AGE_{25-34}$ ); and 3) the proportion of the labor force employed in manufacturing ( $\ln MFG$ ). Table 1 presents these results.

Despite the imperfect specification of such a reduced-form equation, the results are generally consistent with labor-market theory. The productivity index has the proper sign and a coefficient that suggests that a 1 percent increase in productivity is associated

TABLE 1—DETERMINANTS OF DECYCLED LOW-WAGE EMPLOYMENT SHARE YEAR-ROUND FULL-TIME WORKERS: 1963–86

	Coefficient	t-statistic
Constant	11.71	(2.24)
$\ln Q_{perH}$	-2.32	(2.94)
$\ln \%AGE_{25-34}$	.27	(.31)
$\ln \%MFG$	-1.37	(1.89)
$\bar{R}^2$	.43	
SEE	.123	
D.F.	20	
D.W.	.646	

Note: Dep. Var.:  $\ln$  GNP-decycled Low-Wage Employment Share.

with more than a 2 percent decline in the low-wage share. Declines in the manufacturing proportion of total employment have led to increases in low wages, with approximately unit elasticity—a result at least qualitatively consistent with the deindustrialization hypothesis. After controlling for these two factors (and the business cycle), the coefficient on the baby-boom term is not significantly different from zero. In a similar equation, substituting the female share of the employed labor force for the  $AGE_{25-34}$  term, we also failed to find a statistically significant relationship linking the growth in the female workforce to the U turn in wages.

What is most important to recognize, however, is how much is *not* explained by the regression in Table 1. Only about 40 percent of the total variance in the decycled low-wage trend can be explained by these variables. Moreover, the increase in the low-wage share after 1976—controlling for the business cycle, productivity growth, the baby boom, and the relative decline in manufacturing employment—is about as large as the actual decycled trend itself: roughly 5 percentage points. Thus, in spite of falling unemployment rates and rising productivity, the low-wage share continued to increase throughout the 1980's.

#### VI. Conclusion

What are we to make of these results? First, nothing in the data suggest anything but a rising low-wage share and growing

wage polarization, at least after 1979. Second, neither the business cycle nor the entrance of the baby-boom generation into the labor force have contributed more than marginally to the proliferation of low-wage employment (at least among *YRFT* workers). On the other hand, slow productivity growth and the shift of employment out of manufacturing have both played a role in generating the low-wage U turn.

But something else is going on *within* both the service and manufacturing sectors of the economy that is leading to a slippage in wages at the bottom and polarization throughout. Elsewhere, we suggest that this has to do with such factors as the decline in unionization, the erosion in the real value of the minimum wage, the widespread existence of wage concession bargaining, and the institution of two-tier wage structures in a number of large industries. The growing business practice of "outsourcing" to achieve lower labor costs, and the secular shift of capital from directly productive to overtly speculative investment, may also be playing a part in this drama (see our 1988 book). Certainly these broad political economic fac-

tors, together with changes in the relative political power of labor and management during the Reagan era, deserve further investigation.

## REFERENCES

- Bluestone, Barry and Harrison, Bennett, "The Great American Jobs Machine," U.S. Joint Economic Committee, December 1986.
- \_\_\_\_\_ and \_\_\_\_\_, *The Great U-Turn: Corporate Restructuring and the Polarizing of America*, New York: Basic Books, 1988.
- Kosters, Marvin H. and Ross, Murray N., "The Distribution of Earnings and Employment Opportunities: A Re-examination of the Evidence," *AEI Occasional Papers*, September 1987.
- Lawrence, Robert Z., "Sectoral Shifts and the Size of the Middle Class," *The Brookings Review*, Fall 1984, 3-11.
- Levy, Frank S. and Michel, Richard C., "An Economic Bust for the Baby Boom," *Challenge*, March/April 1986, 29, 33-39.
- Norwood, Janet, "The Job Machine Has Not Broken Down," *New York Times*, February 22, 1987.

# The Reemergence of Segmented Labor Market Theory

By WILLIAM T. DICKENS AND KEVIN LANG\*

According to dual labor market theory, the labor market can be usefully described as consisting of two sectors: a high-wage (primary) sector with good working conditions, stable employment, and substantial returns to human capital variables such as education and experience, and a low-wage (secondary) sector with the opposite characteristics. Moreover, primary jobs are rationed, that is, not all workers who are qualified for primary sector jobs and desire one can obtain one. Finally, the sector of the labor market in which an individual is employed directly influences his or her tastes, behavior patterns, and cognitive abilities. Thus the dual labor market model or, more generally, segmented labor market models, is simultaneously a description of the income distribution, a claim about the absence of market clearing, and a radical departure from the standard neoclassical assumption of fully rational actors and exogenously determined preferences. While this last element is potentially the most interesting, even its proponents fail to give it the attention it deserves, and related work has not been incorporated into the segmented labor market model. In this paper, we therefore concentrate on the first two elements of the model.

Segmented labor market theory was sufficiently popular in the late 1960's and early 1970's to be taken seriously by prominent mainstream labor economists. However, two influential and largely negative reviews (Glenn Cain, 1976; Michael Wachter, 1974) portrayed the segmented labor market hypothesis as largely atheoretical and based,

at best, on questionable statistical analysis. It seems fair to say that even sympathetic mainstream critics felt that key insights from the segmented labor model could be incorporated into neoclassical analysis and that the remaining elements of the model did not form a sufficiently coherent theory to pose a challenge to the neoclassical model. Whatever the merits of this perception, it is clear that advocates of the segmented labor market approach did not develop a formal theory which conformed to the standards of mainstream economists. With some notable exceptions (Michael Piore, 1975; David Gordon, 1972), the work was atheoretical. Moreover, the empirical methods used tended to fall outside the norm (for example; interviews, observational studies, and historical and institutional analysis). Advocates of the segmented labor market perspective, mostly radical political economists, chose instead to develop their own research program outside the mainstream.

The reemergence of segmented labor market theory is linked with the reversal of these two tendencies. The theory has been pursued by economists using modern tools of imperfect information theory and state-of-the-art econometrics. As a result, the approach has again attracted the attention of the mainstream. Even a few years ago, it would have been a clairvoyant observer who predicted that Lawrence Summers would be working on a theoretical model of labor market duality (with Jeremy Bulow, 1986), that Robert Solow would count among his recent work a dual market model (with Ian McDonald, 1985) and that James Heckman would publish an article in which he undertook an empirical test of a dual market model, failed to reject the model, and then devoted much of the rest of the article to attacking his and other tests of the dual labor market view (see his article with V. Joseph Hotz, 1986).

Since the theoretical developments are largely associated with efficiency wage and

\*Departments of Economics, University of California, Berkeley, CA 94720 and NBER, and Boston University, 270 Bay State Road, Boston MA 02215 and NBER, respectively. This study was supported in part by NSF grant no. SES-8606139. Lang acknowledges support from a Sloan Faculty Research Fellowship; Dickens acknowledges support from the Institute of Industrial Relations at Berkeley.

related models that are fairly well known and have been reviewed in several places (Lawrence Katz, 1986; Joseph Stiglitz, 1987), in this paper, we concentrate on the empirical work in Section I. In Section II, we discuss theoretical developments briefly and outline important areas for further research.

### **I. Empirical Studies of the Segmented Labor Market**

We have often heard it said that the dual labor market model is untestable (see Heckman and Hotz). The statement that the dual labor market model provides a more parsimonious model of the wage distribution than a single labor market model is eminently testable (see the closely related work of Heckman and Guilherme Sedlacek, 1985). Therefore, the statement that dual labor market theory is untestable is really a statement that it is impossible to test market clearing. If it is truly impossible to test market clearing, then the dual labor market model is untestable, but then so is the human capital model (a position taken by Heckman-Hotz), and it is therefore no more a criticism of the dual labor market model than of neoclassical economics.

There is a sense in which all theories are untestable. However, if we subscribe to this view, "progress" in science is merely a social-psychological phenomenon. We prefer a dynamic view of the selection of models. While no single finding is ever inconsistent with a broad model, models that continue to allow the development of new predictions that are verified are preferable to models that continually must be readjusted to account for new empirical developments. On this basis, the segmented labor market has significantly outperformed the more traditional model over the last few years. We begin with a discussion of the dual labor market model as a description of the wage distribution and then move on to a discussion of market clearing.

#### **A. Wage Distribution**

There are two questions that can be asked about the dual labor market model and the

wage distribution: 1) whether the dual labor market model outperforms a single labor market model; 2) whether the dual labor market model performs well by some absolute standard. The answer to both questions is "yes."

The standard single labor market model estimates a log wage equation with an identically independently and distributed random error term. In effect, the model assumes that the return to education, experience, and other variables is identical for all individuals. This is implausible. Moreover, it is well known that if the error is assumed to be normally distributed, the model provides a poor description of the wage distribution.

In contrast, the dual labor market model of our 1985a article assumes that wages are determined by two log-wage equations (one for each sector). In addition, there is an equation which determines in which sector individuals are employed (the switching equation). Under the assumption that all the error terms are normally distributed, the two wage equations can be estimated jointly with the switching equation, even if sector of employment is not observed by the econometrician. Estimates of the model conform closely to the predictions of the dual labor market model—there is a high-wage sector with much lower returns to these variables. Moreover, since the single labor market model is nested in the dual labor market model, it is possible to test the single labor market model. It can be rejected at any conventional level of significance.

This comparison is somewhat unfair. It is implausible that the log-wage equation is exactly linear and that the error term is homoskedastic. It is possible that the dual labor market model merely improves on the single labor market model by capturing some of this nonlinearity and heteroskedasticity. Consequently, we have submitted the dual and single labor market models to a fairer competition (see our 1987 paper). We developed a single labor market model with higher-order terms in the equations (to allow for nonlinearities) and considerable heteroskedasticity. The number of parameters in the single labor market model exceeded the number in the dual model by one.

Nevertheless, the dual labor market model provided a much better approximation to the empirical wage distribution than did the single labor market model. Formally, using a *chi*-squared goodness of fit test, the single labor market model was easily rejected at conventional significance levels while the test-statistic for the dual labor market model was significant at exactly the .1 level. Moreover, the discrepancies between the empirical distribution and the predictions of the dual labor market model were due almost entirely to spikes in the reported wage distribution at \$7.50 and \$10.00 an hour.

In sum, the dual labor market model not only outperforms an equally complex single labor market model, but it provides a succinct description of the wage distribution. We do not propose that the labor market consists of exactly two distinct segments, only that dualism is a useful simplification. There is a growing literature which emphasizes more general labor market segmentation. Most of this literature uses industry as the basis of analysis. Industry is an imperfect basis for describing segmentation; our analysis indicates that almost all industries have both primary and secondary elements (see our 1985b paper). Nevertheless, it simplifies empirical analysis to equate industries with segments and thus has made investigation of market clearing easier. We turn now to this topic.

### B. *The Failure of Market Clearing*

Our estimates of the dual labor market model cast doubt on workers' ability to choose their sector of employment. Our estimates suggest that most workers would earn considerably more in the primary than in the secondary sector. Unless, workers in the secondary sector have strong preferences for the nonpecuniary characteristics of secondary employment, or the correlation between the error terms in the two wages equations is close to  $-1$ , most workers in the secondary sector must be there involuntarily.

More formally, under certain assumptions described in our earlier papers, the coefficients in the switching equation should be proportional to the difference between the

parameters in the two wage equations. Some departures from proportionality can be justified if workers' characteristics influence the compensating differential they require for secondary work. Our results indicate that either aversion to secondary work decreases with education, blacks are less averse to secondary work than are whites, or that some workers are confined to the secondary sector involuntarily.

While the dual labor market results cast doubt on market clearing, most of the investigation of market clearing has centered on interindustry wage differentials (Dickens and Katz, 1987a, b; Robert Gibbons and Katz, 1987; Jean Helwege, 1987; Alan Krueger and Summers, 1987, forthcoming; Kevin Murphy and Robert Topel, 1987a, b). The results of this series of studies can be summarized as follows. There are large interindustry wage differentials. These differentials are highly correlated over periods of nearly a century so that they cannot be attributed to temporary disequilibria. Therefore, if they are to be explained by a market-clearing model, high-wage industries must either pay a compensating differential or hire unusually skilled workers. Including measures of working conditions seems, if anything, to increase interindustry wage dispersion. That casts doubt on the compensating differentials story. Moreover, the wage differentials are highly correlated across occupations. It is difficult to see why if working conditions were bad for unskilled workers, they would also tend to be bad for secretaries. It is also difficult to see why industries that need unusually skilled laborers also need unusually skilled managers. It is unlikely that there is sufficient complementarity among different types of workers to explain the high level of correlation.

Much of the debate over whether skill differences account for industry wage differentials has focused on whether the wage differentials calculated for a sample of industry-changers is the same as for a cross section of workers. A reasonable summary of recent results is that the former method gives differentials that are highly correlated with the latter, but which may be somewhat smaller in magnitude. A market-clearing

model that is consistent with this result is that some workers have skills that are useful only in high-wage industries. Combined with a model of mobility, this can produce the pattern of differentials among industry-changers. Gibbons and Katz derive some implications of a plausible version of this model. Their preliminary results are inconsistent with the predictions of the model.

A final point consistent with the view that workers in high-wage industries receive rents is the finding that quit rates in these industries are lower. While it is not impossible to derive a market-clearing model with this property, it arises more naturally in a model in which the labor market does not clear.

## II. Unsolved Problems and Potential Applications

At one time, segmentation theory may have suffered from a lack of theoretical foundations—the most difficult problem being the explanation of why wages would remain high in the presence of unemployed workers. Now the problem is choosing among the many competing explanations. It is necessary to do this because the policy implications of market segmentation will differ depending on whether the failure of market clearing is due to principal-agent problems, rent sharing, wage norms, or some combination of causes. For example, if wage differences between sectors are due to principal-agent problems, then there are welfare-improving trade and/or industrial policies that may be pursued to increase high-wage/high-marginal-product employment. But, if wage differences are due to rent sharing, the same policies may affect wages more than employment and act mainly as regressive transfers.

One of the major reasons for the popularity of efficiency wage and related models is that they provide simple and believable explanations for unemployment. However, there is a large gap between the simple models presented in most existing theoretical work and the diversity of experience of unemployed people in the real world. Even in recent models (Bulow-Summers; Stephen Jones, 1985) all unemployment is wait unem-

ployment—workers are unemployed while queuing for primary sector jobs. In contrast, earlier work on segmented labor market (Piore), emphasizes flux and uncertainty in the secondary sector as the source of unemployment.

In fact, both sources of unemployment arise naturally in an efficiency wage model of labor market segmentation. If there is a sufficient advantage to being unemployed while searching, some workers will prefer to remain unemployed while seeking high-wage employment. So there will be some wait unemployment. Layoff from the primary sector will be likely to result in wait unemployment since such workers will have more unemployment insurance and accumulated assets which will make waiting more affordable.

If primary firms pay high wages to reduce shirking, stealing, absenteeism, quits, etc., they will find it advantageous to provide stable employment since the longer workers' horizons, the lower the wages required to deter such behavior. Alternatively, if primary firms are capital intensive, they may attempt to keep production stable to insure that the capital, and consequently the workers, are steadily employed. Secondary firms that do not pay high wages may well have a cost advantage in serving unstable demand. Workers in the secondary sector would suffer considerable frictional unemployment as the market adjusted to continual shifts in the level and distribution of demand.

One of the remaining challenges for segmented labor market theorists is to develop such a model more fully and to subject it to empirical scrutiny. In our opinion, however, such a model still suffers from the failure to consider some of the more radical elements of the early segmented labor market literature. In particular, the scarring effect of secondary employment is worthy of further investigation.

## III. Conclusions

The failure of the market-clearing assumption has enormous significance not only for labor economics, but also for such fields as

macroeconomics and international trade. Prescriptions based on the assumption of market clearing may be grossly inaccurate when markets do not clear. At the same time, we must exercise caution. We have yet to derive a fully articulated and satisfactory model of labor market segmentation, and policies based on tentative models must be viewed with skepticism. Some efficiency wage models imply that it would be desirable to raise the cost to workers of being unemployed. This might be a terrible injustice if that form of the model were inappropriate.

Despite these caveats, it is clear that work on segmented labor markets has made dramatic strides in the last few years. As a description of the wage distribution, dual labor market models are significant improvements over the more standard one-sector models and, given data and theoretical limitations, provide a surprisingly accurate description of the empirical distribution. The evidence against market clearing is accumulating so that it is becoming increasingly necessary to develop more fully articulated models consistent with these facts.

## REFERENCES

- Bulow, Jeremy I. and Summers, Lawrence H., "A Theory of Dual Labor Markets with Application to Industrial Policy, Discrimination and Keynesian Unemployment," *Journal of Labor Economics*, June 1986, 4, 376-414.
- Cain, Glenn, "The Challenge of Segmented Labor Market Theories to Orthodox Theory," *Journal of Economic Literature*, December 1976, 14, 1215-57.
- Dickens, William T. and Katz, Lawrence F., (1987a) "Interindustry Wage Differences and Industry Characteristics," in Kevin Lang and Jonathan S. Leonard, eds., *Unemployment and the Structure of Labor Markets*, Oxford: Basil Blackwell, 1987.
- \_\_\_\_\_ and \_\_\_\_\_, (1987b) "Interindustry Wage Differences and Theories of Wage Determination," NBER Working Paper, 1987.
- \_\_\_\_\_ and Lang, Kevin, (1985a) "A Test of Dual Labor Market Theory," *American Economic Review*, September 1985, 75, 792-805.
- \_\_\_\_\_ and \_\_\_\_\_, (1985b) "Testing Dual Labor Market Theory: A Reconsideration of the Evidence," NBER Working Paper, 1985.
- \_\_\_\_\_ and \_\_\_\_\_, "A Goodness of Fit Test of Dual Labor Market Theory," NBER Working Paper No. 2350, 1987.
- Gibbons, Robert and Katz, Lawrence, "Learning, Mobility, and Inter-Industry Wage Differences," unpublished, MIT, 1987.
- Gordon, David M., *Theories of Poverty and Underemployment*, Lexington: D.C. Heath, 1972.
- Heckman, James J. and Hotz, V. Joseph, "An Investigation of the Labor Market Earnings of Panamanian Males: Evaluating the Sources of Inequality," *Journal of Human Resource*, Fall 1986, 21, 507-42.
- \_\_\_\_\_ and Sedlacek, Guilherme, "Heterogeneity, Aggregation and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market," *Journal of Political Economy*, December 1985, 93, 1077-125.
- Helwege, Jean, "Interindustry Wage Differentials," unpublished, UCLA, 1987.
- Jones, Stephen R. G., "Minimum Wage Legislation in a Dual Labor Market," unpublished, University of British Columbia, 1985.
- Katz, Lawrence F., "Efficiency Wage Theories: A Partial Evaluation," *NBER Macroeconomics Annual 1986*, Cambridge: MIT Press, 1986, 235-76.
- Krueger, Alan and Summers, Lawrence H., "Efficiency Wages and the Inter-Industry Wage Structure," *Econometric*, forthcoming.
- \_\_\_\_\_ and \_\_\_\_\_, "Reflections on the Inter-Industry Wage Structure," in Kevin Lang and Jonathan S. Leonard, eds., *Unemployment and the Structure of Labor Markets*, Oxford: Basil Blackwell, 1987.
- McDonald, Ian M. and Solow, Robert M., "Wages and Employment in a Segmented Labor Market," *Quarterly Journal of Economics*, November 1985, 100, 1115-41.
- Murphy, Kevin M. and Topel, Robert H., (1987a) "Unemployment, Risk, and Earnings," in

- Kevin Lang and Jonathan S. Leonard, eds., *Unemployment and the Structure of Labor Markets*, Oxford: Basil Blackwell, 1987.
- \_\_\_\_\_ and \_\_\_\_\_, (1987b) "Efficiency Wages Reconsidered: Theory and Evidence," unpublished, University of Chicago, 1987.
- Piore, Michael J., "Notes for a Theory of Labor Market Segmentation," in Richard C. Edwards et al., eds., *Labor Market Segmentation*, Lexington: D. C. Heath, 1975.
- Stiglitz, Joseph E., "The Causes and Consequences of the Dependence of Quality on Price," *Journal of Economic Literature*, March 1987, 25, 1-48.
- Wachter, Michael L., "Primary and Secondary Labor Markets: A Critique of the Dual Approach," *Brookings Papers on Economic Activity*, 2:1974, 637-80.



## Contractarian Political Economy and Constitutional Interpretation

By JAMES M. BUCHANAN\*

At the meetings in 1974, I presented a paper entitled, "A Contractarian Paradigm for Applying Economic Theory" (printed 1975). In that paper, along with others (see my 1979 book), I argued that our subject matter is centrally a "science of exchange" or a "science of contract," and that the exchange paradigm should take precedence over the maximizing paradigm. This shift in the focus of positive inquiry carries normative implications. Conceptions such as aggregate efficiency in the allocation of resources become, at best, examples of functionalist error, along with the more explicitly normative variants of the social welfare function. The contractarian or catallactic approach to economic interaction suggests that systems or subsystems be evaluated in terms of the comparative ease or facility with which voluntary exchanges, contracts, or trades may be arranged between and among members of the community. Normative judgments take the form of statements that array "better" and "worse" *processes* (rules, laws, institutions) within which exchanges are allowed to take place. These judgments are categorically distinct from those that array and evaluate results or outcomes.

This shift in normative political economy has implications for the issues of constitutional interpretation debated by legal scholars and philosophers. These issues involve disputes along several related and intersecting dimensions: between judicial activism

and nonactivism; between judicial deference to legislative authority and judicial independence; between strict constructivism and pragmatism; between original intent and legal environmentalism; between teleological and deontological conceptions of law. My purpose here is to discuss some of these contractarian implications for constitutional interpretation. This is a limited purpose, and I advance no direct and extended argument on either general philosophical issues, or on points of debate in particular legal settings. Any identifiable contribution of the contractarian political economist must emerge from the differentially abstracted order that his perspective imposes on social reality.

Section I covers the familiar distinction between an individualistic and a communitarian starting point. The implication for legal interpretation of constitutional rules is almost self-evident. In Section II, I again go over analysis, developed elsewhere, that extends the catallactic paradigm from the economy to the political order, and, in particular, to the design, selection, and enforcement of constitutional rules. Section III examines the implications for judicial interpretation of the political constitution, and, in particular, the implications for the debate between strict constructivism and pragmatism and between original intent and legal environmentalism. In Section IV, the argument is extended to the contractarian's stance in interpretative confrontation with rules that cannot find a logic in any contractarian ideal.

### I. Normative Individualism

The primary question in any contractarian perspective on social order is the definition

<sup>†</sup>*Discussants:* Richard Wagner, Florida State University; Nathan Rosenberg, Stanford University.

\*Center for Study of Public Choice, St. George's Hall, George Mason University, Fairfax, VA 22030.

of the units that potentially engage in exchange. The economists' response here is straightforward; individuals enter into exchange, one with another, either to make direct trades of goods and services, or to create organizations (firms, clubs, states, associations) that, in turn, make such trades on their behalf. If the community exists as an organic entity in some sense prior to and independently of its individual members, and, further, if this community has its own supra-individualistic goals, the exchange perspective clearly breaks down. With whom could the inclusive community, as such, make exchanges?

If, however, the organic or communitarian paradigm is rejected in favor of an individualistic one, implications emerge that embody both methodological and normative content. If individuals, or organizations of individuals, are the units that enter into exchanges, the values or interests of individuals are the only values that matter for the quite simple reason that these are the only values that exist. Such terms as "national goals," "national interest," and "social objectives" are confusing at best. Individuals in a community may, of course, share values in common and they may agree widely on specific goals or objectives for policy directions to be taken by their political organization. But this very organization, like others, exists only for the purpose of furthering individual values and interests.

This summary of the normative individualist's position is sufficient to suggest the direct implications for constitutional interpretation. The "good society" is that which best furthers the interests of its individual members, as expressed by these members, rather than that society that best furthers some independently defined criterion for the "good." The basic "rules of the game," the law, cannot be conceived as a means through which the community is shifted toward that which judges or intellectuals deem to be good. Any teleological conception of the law, and of the constitution and of the role of the judiciary, is simply out of bounds under any contractarian or exchange conceptualization of social order.

## II. Political Exchange

If we adhere strictly to the individualistic benchmark, there can be no fundamental distinction between economics and politics or, more generally, between the economy and the polity. The state, as any other collective organization, is created by individuals, and the state acts on behalf of individuals. Politics, in this individualistic framework, becomes a complex exchange process, in which individuals seek to accomplish purposes collectively that they cannot accomplish noncollectively or privately in any tolerably efficient manner. The catallactic perspective on simple exchange of economic goods merges into the contractarian perspective on politics and political order.

But how can ordinary politics as we observe it possibly be modeled as a complex exchange process in which individuals *voluntarily* participate, at least in any sense at all analogous to their participation in markets? Any attempt to extend the exchange perspective to politics seems absurd on its face, since we observe politics to be characterized by conflict rather than cooperation, best modeled as a game that is zero or negative sum. Coercion rather than voluntary participation seems to be the primary relationship embodied in politics. If, however, this coercion-conflict element is elevated to center stage, how can the state ever be legitimized or justified to the individual?

A way out of the apparent paradox is provided if we shift attention from ordinary politics, which is almost necessarily majoritarian, or, more generally, nonconsensual in its operation, to constitutional politics, which may at least approach consensual agreement, at least in its idealization. Individuals may generally agree upon the rules of the game within which ordinary politics takes place, and these agreed-on rules may allow for predicted net gainers and net losers in particularized political choices. The question of legitimacy or justification shifts directly to the rules, to the constitutional structure, which must remain categorically distinct from the operations of ordinary politics, which is constrained by the rules. As noted

earlier, the argument does have direct implications for judicial interpretation. The most critical of these implications stems from the categorical separation itself. There is a critically important functional role for judicial review. The "state-as-umpire" function is properly assigned to a branch of the political order that is separated from those branches that operate within the rules. Further, this function is conceptually as well as operationally different from ordinary politics. The judiciary, in its umpire role, must take a truth-judgement approach, an approach that is inappropriate in the workings of ordinary politics. The judiciary must determine whether or not the rules have been violated, whether or not a rule exists, whether or not a rule applies to this or that case. These are truth judgments. It becomes absurd to introduce arguments based on such things as "compromises among interests" or "proper representation of interests" in the whole judicial exercise.

### III. Changes in Rules

In an earlier paper (1986), I classified the inclusive political order in terms of three separate functions. The first involves the enforcement of the rules that exist. This embodies the role for judicial review that I have just discussed. The second involves the carrying out of ordinary politics within the rules that exist. This includes taxing, spending, and other activities within the broad rubric summarized as the financing and supply of public goods and services. The third function involves changes in the rules themselves, or constitutional reform. I have argued above that the judiciary, as an independent branch of the political order, properly operates within the first of these three functions. The legislative body, reflecting the interplay of groups interests, properly operates within the second of these functions. The third function, that of changing the rules, is inappropriate both for the judiciary and for the legislative branch. The rules are changed only through the well-defined procedures for constitutional amendment, procedures that are explicitly more inclusive than ordinary legis-

lation or judicial review. A straightforward implication of the contractarian complex exchange perspective on political order is that the judiciary oversteps its proper limits when it takes on the task of changing the basic rules within which the socioeconomic-legal game is played.

If the judicial function is, and must be, restricted to interpretation of the rules that exist, specific guidelines for judges and courts charged with constitutional interpretation necessarily emerge. Parallel to this restriction on the scope of the judicial role is that placed on the legislative role. The legislature also oversteps its proper limits when it moves beyond existing boundaries and itself makes changes in the constitutional order. From this it follows directly that the judiciary should not be deferential to legislative decisions when these have the effect of modifying the basic rules. In this respect, an activist, rather than a nonactivist, and deferential court is required.

However, because the judicial role is itself limited to interpreting rules that exist, and cannot go beyond this, something akin to strict constructivism seems to be implied here. But the important point to be made is that the court should act as strict constructivist with respect to the constitutional rules that exist in the status quo when the case at issue is confronted. This status quo may, but need not, reflect generally accepted rules that are readily derivative from the original intent of those who designed and emplaced the written documents. The rules reflecting original intent may have been gradually modified by the historical case record to the point where there seems little connection with the rules that describe the status quo.

The indeterminacy in defining the status quo is unsatisfactory to many strict constructivists. How can a court define the rules that exist without direct resort to something like original intent? It is here that the court needs to rely on something akin to the modern economists' notion of rational expectations. Those rules in existence are those that best describe the set of individuals' expectations about the boundaries of political

authority when the activities in question were carried out. An analogy from ordinary games may be helpful here. Suppose that the rule book describing the activities that may take place within a game, say, basketball, has remained unchanged for a number of years. But as the game has evolved, within the changing technology and changing skill levels of players, referees have gradually and incrementally modified the effective rules, for example, on walking with the ball. A new referee fulfills his role properly when he tries to enforce the rules that exist; he violates his assigned task when he tries to go back and enforce strictly the rules-as-written in the outdated rule book.

#### IV. "Bad" Rules

The argument to this point is noncontroversial in the sense that the suggested implications for judicial interpretation of the constitution follow straightforwardly from acceptance of the individualistic-contractarian perspective on political order. A more debatable set of issues arises as we focus attention on the stance of the judge, who fully shares the contractarian perspective, who is faced with the status quo existence of rules that reflect neither original intent nor plausibly justifiable extensions of such intent, and that, further, could never have passed any conceivable contractarian consensus test for legitimacy, even in some conceptual sense. That is to say, there may exist rules which are contained within the expectational set of both citizens and ordinary politicians, that have been imposed nonconsensually. Should the contractarian judge move beyond mere enforcement of the status quo in some attempt to dismantle "bad" law?

Much modern economic regulation (for example, minimum wage laws, rent control laws) presumably fits this category. Should the contractarian constitutionalist deem such laws to be nonconstitutional, despite the fact that prior courts have made judgements to the contrary? My argument suggests that if the prior judicial interpretations have been in place sufficiently long for these interpretations to have formed part of the rational

expectations of both the citizenry and the acting political agents, it would not be appropriate for the contractarian judge to seek actively to change the rules. In this respect, my argument places me squarely on the Scalia side of the Scalia-Epstein debate. (Antonin Scalia, 1987; Richard Epstein, 1987.) Retrospectively, the court must defer to the status quo set of rules that exist, which may well embody prior judicial approval for legislation (including judicial legislation) that unconstitutionally shifted the boundaries of the consensual order. To move beyond such deference to the status quo and to assume an activist role in deconstruction, as guided by some ideal, even if this ideal be contractarian, opens up judicial review to precisely those dangers of abuse that Scalia warns against.

On the other hand, my argument suggests that the contractarian judge should be quite jealous in his protection of the existing rules from legislation and judicial intrusion that fails the consensual test. In a prospective or *ex ante* sense, my argument places me on the Epstein side of the debate with Scalia. Deference to legislative authority, per se, cannot be justifiably derived from the three-stage contractarian model of political order outlined.

Scalia argues that the courts should remain passive as legislatures act to constrain economic liberties. Epstein argues that courts should act to protect economic liberties, whether the legislation constraining such liberties has long existed and been upheld by prior court judgments or whether such legislation is recent or newly proposed. Neither Scalia nor Epstein makes the temporal cut that my argument implies. The contractarian position, as I interpret it, requires that the rules that exist, no matter how these might have come into being, be treated as relatively absolute absolutes and enforced by the courts until and unless these rules are changed by defined procedures for change.

My position does depend critically on some ability to define meaningfully just what the set of status quo rules is, an ability that is not centrally important to either Scalia or Epstein. And in this respect I return to the importance of the expectational setting, to

which courts should remain highly sensitive. Any legislatively orchestrated change that upsets the legitimately held expectations of citizens should be interpreted as a change in the constitutional structure, and, as such, should be prevented by the courts.

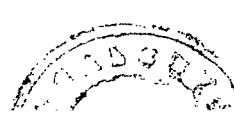
Consider the much discussed taking of property for public purpose. Modern courts have allowed legislatures authority to modify values of privately owned property within very broad public purpose limits. But there do remain limits, and wholly arbitrary intrusion would, presumably, be rejected even by modern courts. My position suggests that courts carefully draw such limits at the set of expectations held in the status quo, as properly measured.

In the three-stage functional classification imposed on political order by the contractarian perspective, the role for the judiciary is clear. The function of the judiciary is protection of that which is, which remains perhaps the most critical function for the maintenance of order and stability. The judicial branch properly serves a stabilizing rather than a reformist or restorationist role. The courts should protect what is rather

than try to promote what might be, or try to restore what might have been.

## REFERENCES

- Buchanan, James M., "A Contractarian Perspective for Applying Economic Theory," *American Economic Review Proceedings*, May 1975, 65, 225-230.
- \_\_\_\_\_, *What Should Economists Do?*, Indianapolis: Liberty Press, 1979.
- \_\_\_\_\_, *Liberty, Market, and State*, New York: New York University Press, 1986.
- \_\_\_\_\_, and Tullock, Gordon, *The Calculus of Consent*, Ann Arbor: University of Michigan Press, 1962.
- Epstein, Richard, "Judicial Power: Reckoning on Two Kinds of Error," in J. Dorn and H. Manne, eds., *Economic Liberties and the Judiciary*, Fairfax: George Mason University Press, 1987, 39-46.
- Scalia, Antonin, "Economic Affairs as Human Affairs," in J. Dorn and H. Manne, eds., *Economic Liberties and the Judiciary*, Fairfax: George Mason University Press, 1987, 31-37.



# Original Intent, History, and Doctrine: The Constitution and Economic Liberty

By HARRY N. SCHEIBER\*

One of the enduring and truly perplexing issues in American legal theory is the question of how and to what degree the values associated with economic liberty are embodied in the United States Constitution. Recent controversies, both in law and in economics, on the efficiency of the public sector in regulation and in the provision of public goods, on the virtues or perils of privatization, and on public choice and the legal process proceed too often without reference to the American system's constitutional heritage. The assumptions often manifest in neoconservative analysis, particularly the view that "economic liberty" should be understood exclusively in *laissez-faire* terms (see, for example, Richard Epstein, 1986), need to be considered not only with regard to their theoretical merits and prescriptive implications, but also with regard to an accurate understanding of their historical context.

Hence this paper reexamines the nexus between the Constitution and the idea of economic liberty. The inquiry proceeds on three lines. First, it deals with how concepts of economic liberty connect with "original intent." To what degree did the Constitution of 1787—as modified by subsequent amendments whose vital importance is too often overlooked by scholars who argue from the "original understanding"—and its commitment to a government of limited powers embrace the idea of strict restraints upon economic interventions that would abridge private-sector liberties?

Second, the various historic meanings of economic liberty are examined in the context of evolving constitutional doctrines. This part of the inquiry cannot be bounded by the formal decisions of the U.S. Supreme Court; rather, it must embrace the history of state policies and law. The broader context is required because in vital respects—including not only a broad range of positive governmental interventions but also the ordering of basic rules of property relationships—it was the state governments, and only marginally the federal government and constitutional law, which (as shown in my 1987 paper) played the central role.

Third, the paper considers an emerging view in studies of constitutional theory that has reconceptualized economic liberty as a jurisprudential concept that fully embraces values of individual autonomy often running contrary to the traditional concept of private-sector claims to vested rights in property. Whatever their merits or failures, constitutional theories of autonomy linked to economic liberty as a constitutional value—like modern theories of public-choice behavior—highlight the need for an accurate understanding of the place that has been given historically to various concepts of intervention, redistribution, and property protection in American constitutional and legal history.

## I

Given eighteenth-century political and jurisprudential values in the Anglo-American world, it seems an irrefutable premise that the values associated with economic liberty were intended to be advanced in significant ways by the Constitution of 1787, including the Fifth Amendment on takings and on due process. Analysis of original intent would be much easier, to be sure, if we could simply equate economic liberty

\*Professor of Law, Boalt Hall School of Law, University of California, Berkeley, CA 94720. I am indebted for valuable comments to J. R. T. Hughes, Thomas Jorde, Kent Newmyer, Jane L. Scheiber, and David Thelen.

with the security of vested rights in property; for several important provisions of the Constitution express doctrines that complement or are proxies for the common-law guarantees of the right to quiet possession and of a property owner's reasonable expectations.

Even the most dogmatic defenders of property rights in the founding era, however, recognized that property was in legal terms a bundle not only of *rights* but also of *obligations*—owners being subject to taxation, regulation in the public interest, and even physical taking by power of eminent domain or emergency seizure. The most conservative jurists of the early republic—Justices Patterson, Marshall, Story, and Chancellor Kent—recognized fully the state's power of eminent domain, while also recognizing a broad power in the state to tax, to channel economic activity and institutional development by giving rights in one kind of property priority over other rights, and to regulate through the police power for purposes of advancing health, safety, and welfare (James Hurst, 1977). A redistributive function, trenching on property rights, thus was recognized from the outset as wholly legitimate. Common-law and constitutional doctrines, as well as statutes, were mobilized not only to serve the values of privatism—the vested rights of private owners—but (as shown in Jonathan Hughes, 1977; Hurst; myself, 1984) also to serve communal or societal values and interests, usually in overt tension with the values of privatism and economic individualism.

Historical analysis of “economic liberty” in relation to property law and constitutional values more generally therefore must take into account the possibility that “liberty” has meant something much more than crude Lockean precepts of economic individualism would suggest. It must acknowledge also the communal strain in the American legal and constitutional tradition—even in property law—which has qualified and constrained private claims to vested rights.

Indeed, the state courts in the nineteenth century developed a doctrine of “public rights” that was mobilized in the jurisprudence of eminent domain powers, in the

doctrine of public trust (concerning public property so important to the community that its disposition by the state came under special constraints), the police power, and in taxation law. This public rights doctrine was driven by the concept of the public good, or the “commonwealth” ideal; it was invoked to validate a wide range of interventionist policies, as the states responded creatively with a variety of positive programs and regulatory regimes to what Chancellor Kent called the “growing wants and prosperity” of a society under conditions of rapid technological change. State courts built this public rights doctrine upon concepts of public nuisance from the common law; upon the Anglo-American concept of the representative legislature holding the ultimate powers of sovereignty; and upon an essentially materialist view of the imperatives of progress. (See my 1984 and 1987 papers.)

Public rights doctrine was no less expressive of the values of economic liberty than were contemporary federal judicial doctrines under the Contract Clause or property-law doctrines that erected the legal bulwarks of vested rights. The contemporary view of economic liberty, withal, contained elements of competing traditions—the individualistic and atomistic concept of property rights held against government itself and against competitive invasion by other private owners; and the alternative tradition in law which reflected the idea of communal interests—what Chief Justice Roger Taney in the Charles River Bridge decision termed “the comforts, convenience, and prosperity of the people.” In effect, economic advancement and prosperity of the whole polity became one validating legal principle among several; but it was a principle that stood as a variant of economic liberty defined as the whole community's.

The jurisprudence of public rights complemented and reinforced directly a range of interventionist state-level policies that promoted and built transportation systems and other infrastructure; fostered development of corporations, including many that mixed public and private capital; and perpetuated, with market regulations and in other ways, the mercantilist traditions which were also a

significant inheritance in economic thought and policy from the constitutional founding era.

In light of this history, it seems clear, doctrines that permitted state actions which abridged private vested rights and fostered significant redistribution of wealth were a fully integral component of the economic liberty concept. Any claims in theoretical or prescriptive analysis today for a more restrictive, individualistic notion of economic liberty—one that is cast exclusively in terms of stabilizing and protecting private interests—constitutes, in fact, a departure from the historic tradition of the nineteenth century and even from the “original intent” of 1787–91.

## II

Adoption of the Fourteenth Amendment in 1868 opened a new era in the history of economic liberty as a constitutional precept. Whereas legal-constitutional doctrines in the antebellum era had found room for both individualistic/privatistic and communal variants of economic-liberty values, now an express egalitarian ideal was introduced through the twin requirements of due process and equal protection. Ironically, a conservative, property-minded Supreme Court in the era culminating in *Lochner v. New York* in 1905, applied the Fourteenth Amendment in ways that denied equality to blacks while developing a highly privatistic version of property rights that could be used to invalidate social and economic regulatory legislation hostile to corporate and industrial interests (Charles McCurdy, 1975).

Still, however, there occurred an interpenetration of egalitarian legal concepts with economic liberty, first manifest in radical thought of the abolitionists and then embodied in civil rights legislation of the 1860's that preceded adoption of the Fourteenth Amendment itself (for example, the requirement of the 1866 Civil Rights Act that all citizens regardless of race or color might hold and dispose of property on equal terms).

Even while betraying the former slaves who were the intended beneficiaries of this enlarged version of economic liberty as a constitutional ideal, conservative jurists pur-

posefully enlarged the liberty concept by elaborating a doctrine of the “right to pursue a lawful calling,” founding it in the Fourteenth Amendment’s guarantees—and also in the language of the Declaration of Independence, which they imported into the Constitution as part of this elaborate doctrinal maneuver. Although the purpose was conservative, the rhetoric was not: indeed, as McCurdy reminds us, Abraham Lincoln had spoken in defending his Civil War policy of giving all persons “an unfettered start and a fair chance in the race of life.” In this sense, the notion of freedom for all persons to use their faculties to the fullest, in pursuit of their aspirations, individual potential, and “happiness,” in the economic sphere as in others—however effectively the *Lochner* Court may have narrowed and distorted it—was perpetuated in a conservative age.

## III

The conservative jurisprudence of *Lochner* came apart on the rocks of economic crisis and a revival of communal values in the era of the Great Depression and the New Deal. The doctrines of property rights under constitutional protection, expanded so dramatically since the 1880s, were renounced in several leading decisions of the Court. In *Home Building and Loan Association v. Blaisdell* (1934), Chief Justice Hughes, writing for the majority, declared that the economic emergency warranted a state moratorium law affecting mortgage debt obligations. Meanwhile, in *Nebbia v. New York* (1934), the Court abandoned the doctrine of “business affected with a public interest,” which for half a century had been used to overturn regulatory legislation.

In the *Carolene Products Case* (1938), the Court introduced a new double standard for judicial review of regulatory laws. The burden of this double standard was that the Court, in reviewing such laws, would conduct a “more searching judicial inquiry” on behalf of basic political freedoms such as the right of assembly than it would conduct when legislation seemed only to touch rights in property or other economic relationships. Thereafter, the Court steadily elaborated this distinction between “personal rights” and



(mere) "property rights"—never explaining fully, however, why the rights of an individual owner in property were not personal rights as well; the justices seem to have relegated to the "ash-bin of history" property rights in the individualistic or Lockean mode (James Oakes, 1981, p. 608).

Whatever its logical or equitable failures, the double standard embodied in the *Carolene* formula permitted legislative will—the awesome prerogative power—to trump the rights historically associated with private claims to physical holdings or franchises. Also subordinated now, however, to governmental action, in the name of the "public interest," was the entire range of economic rights associated with individualism, including the Fieldian idea of full play for an individual's faculties and potential in a society based on equal rights.

For many years, economic liberty in its individualistic aspect, as a right associated with the Constitution, thus faded from the forefront of attention in academic or legal discourse. The matter of *economic opportunity* remained on the national agenda, but in a policy context and not in doctrinal constitutional terms—relating to programs for welfare and relief, educational opportunity, and the like.

The liberal position, associated with the New Deal and the *Carolene* doctrine that dichotomized "personal" and "economic" rights, had a strongly paternalistic aspect, expressed well in Bernard Schwartz's view, that "in a very real sense, the true liberty of the individual may be promoted by restrictions that the society imposes upon him in his own interest" (1964, p. 205). This view, that expressed faith in survival of individuality within a network of interventionist welfare and regulatory policies, would come under a sustained attack from many fronts—an attack that has established a revised intellectual context for the modern debate over constitutional issues of economic liberty.

#### IV

The attack against paternalism, on the one side, and against the preferred freedoms doctrine, on the other, came in part from jurists such as Justice Potter Stewart, who in

*Lynch v. Household Finance Corp.* (1972) insisted upon "a fundamental interdependence between the personal right to liberty and the personal right in property." A sustained critique also came from scholars such as Hughes, whose studies faulted interventionism and giantism in modern government for its inefficiency and paternalism. Hughes' work did not advance wholesale an argument for minimalism, though it did reflect some of the public-choice critique of interventionism; but other scholars in law and economics went further, making efficiency values and the maximization of social wealth the key criteria for assessing the legitimacy and efficacy of government's role.

From the emergent New Left, meanwhile, came arguments such as those of Charles Reich (1964), whose studies reestablished in constitutional scholarship the once-standard nexus of property and economic liberty. Economic liberty in its individualistic aspect, defined as a basic right, also came back into play in connection with civil rights litigation and debate of the 1960's Civil Rights Acts, which sought to establish for blacks (and, later, other minorities and women) equality of legal status.

In the Supreme Court, in decisions on the right to travel, personal privacy, and entitlement to welfare benefits, a new Fourteenth Amendment jurisprudence was built upon the doctrines of *Brown v. Board of Education* to expand the dimensions of personal "autonomy" as a value in constitutional doctrine. The high-water mark was *Shapiro v. Thompson* (1969), striking down residency requirements for welfare eligibility; here the Court spoke in positive terms of economic liberty, portraying the dream of economic opportunity (not the dole) as the touchstone. The Court was divided sharply, however, in later cases, including two 1987 land-use regulation cases from California in which the Court at least nominally extended important new protection to landowners' rights portrayed in robust Blackstonean terms.

The courts doubtless will continue to struggle with questions of both procedure and substance bearing on issues of property ownership and economic liberty. They do so, however, in the context of a complex doctrinal history that has included several distinct

approaches to the meaning of such liberty in constitutional terms. There is a tradition of public obligations and of public and communal rights, alongside that of private vested rights; and there is a line of doctrine that portrays economic liberty in positive terms, as requiring the state to provide individuals with opportunity. The related concept of "personal autonomy" is thus increasingly recognized formally in constitutional jurisprudence as embracing a set of rights, including entitlements associated with economic liberty.

Neither a unidimensional "vested rights" orthodoxy nor a simplistic dichotomization of personal and property rights can therefore accurately represent the rich complexity of the economic liberty concept in our constitutional law.

#### REFERENCES

- Epstein, Richard A., *Takings: Private Property and the Power of Eminent Domain*, Cambridge: Harvard University Press, 1986.
- Hughes, Jonathan R. T., *The Governmental Habit: Economic Controls from Colonial Times to the Present*, New York: Basic Books, 1977.
- Hurst, James Willard, *Law and Social Order in the United States*, Ithaca; London: Cornell University Press, 1977.
- McCurdy, Charles W., "Justice Field and the Jurisprudence of Government-Business Relations: Some Parameters of Laissez Faire Constitutionalism, 1863-1897," *Journal of American History*, March 1975, 61, 970-1005.
- Oakes, James L., "Property Rights in Constitutional Analysis Today," *Washington Law Review*, November 1981, 56, 583-626.
- Reich, Charles, "The New Property," *Yale Law Journal*, April 1964, 73, 733-87.
- Scheiber, Harry N., "Public Rights and the Rule of Law in American Legal History," *California Law Review*, June 1984, 72, 217-51.
- \_\_\_\_\_, "State Law and 'Industrial Policy' in American Development, 1790-1987," *California Law Review*, January 1987, 75, 415-44.
- Schwartz, Bernard, *A Commentary on the Constitution of the United States, Part II: The Rights of Property*, New York: Macmillan, 1965.

# Contested Exchange: Political Economy and Modern Economic Theory

By SAMUEL BOWLES AND HERBERT GINTIS\*

Contemporary liberal writers share a fundamental assumption: societal rule-making concerns not the political structure of the economy, but rather the political structure of the state. As we shall see, an outmoded variant of microeconomic theory has played no small part in suppressing the acute question of economic power and its accountability.

Concerns of class and power, which have occupied a major position in the thought of Marxists, institutionalists, and historians, thus find no echo in modern liberal orthodoxy. Indeed, in the twentieth century we have seen the ascendancy of a social theory that makes politics and the state co-extensive. This is evident in the identification of the public sphere of society with relations of citizens to the state and the private sphere with familial and contractual relationships; the social relationships in the public sphere are deemed political and the just realm of democratic claims, while the remaining social relations, called private, are seen as lying beyond the limits of political discourse. Thus what is transparent to the unschooled—that the capitalist economy confers power upon those who occupy leading positions in the business world—is denied by the sophisticated.

Neoclassical economics supplies an elegant argument precisely to this effect. The only form of power that this theory recognizes is the command over goods and services, called purchasing power, that is summarized by the location of the relevant budget constraint. The price-taking economic agent enshrined in our texts has infinite power over quantities and none over price. In competitive

equilibrium all markets clear, no agent is quantity constrained, and as a result, all agents are equally powerful or powerless as the case may be. Even the apparent power of the boss is an illusion, for as Schumpeter pointed out long ago, competitive pressures will force bosses (or one might add worker-elected factory councils) to do whatever is cost minimizing given available technologies and factor prices.

Our claim that the capitalist economy has a political structure, one which is undemocratic and on democratic grounds ought to be replaced by a set of rules assuring both liberty and popular accountability of power, will thus strike those with the benefit of training in economics as not so much wrong as nonsensical. Not by ideological inclination alone will most contemporary economists be more at home discussing the Pareto-inferior properties of the U.S. Department of Transportation than the undemocratic political structure of General Motors.

But the silence of economists on the political structure of the economy is unwarranted; perhaps surprisingly, a consistent application of the axioms of rational self-interested individual action does not support the neo-classical view of the apolitical economy but rather, as we shall see, provides the microeconomic basis for the study of economic power.

## I. Contested Exchange: Political Structure and Competitive Markets

By the political structure of the economy, we mean the ensemble of rules governing investment, production, and distribution in economic institutions. An actor who holds decision-making authority within the political structure of the economy is said to have *economic power*. We call decision-making

\*Professors of Economics, University of Massachusetts, Amherst, MA 01003.

authority *substantive* if its exercise has non-trivial allocative or distributive effects.

The political theory implicit in the neoclassical model may then be neatly summarized: in the competitive equilibrium of a capitalist economy, where there is power (in the firm), authority is nonsubstantive and where decision-making authority is substantive (in markets), there is no power.

This treatment of economic power in neoclassical theory flows from three central commitments. The first, which we may term the *neutrality of property assignments* asserts that the distribution of ownership rights has no substantive allocative implications: any Pareto-efficient equilibrium can be supported by some initial assignment of property titles followed by competitive exchange; its most celebrated version is the Coase Theorem, according to which, even in the case of market failures, the initial assignments of titles to property among competitive agents has purely distributional effects in equilibrium.

The second commitment is the *irrelevance of command*, according to which the political structure of firms has neither allocative nor distributive effects in competitive equilibrium. Organizational forms may be more or less efficient, of course, but among efficient forms, the source of decision-making authority or, more formally, the location of sovereignty, is irrelevant: how the quarterback is selected does not alter the way the game is played. The third, the *efficiency of market exchange*, asserts that in the absence of market failures, a competitive market equilibrium is Pareto efficient.

Taken together, these three neoclassical propositions imply a radical separation of distributional, political, and allocational concerns. Were these propositions correct, a wide range of normative objectives, whether egalitarian or democratic, could be implemented simply through a reassignment of either property titles or decision-making rights without direct intervention in the allocational process and, indeed, without allocational effects. Thus while Paul Samuelson's commonplace remark that in a competitive model it makes no difference whether capital hires labor or the other way might

tax the credulity of sophisticated students of social life, it raises no eyebrows among contemporary neoclassical economists—it should.

The still widely accepted model by which the apparent power of economic elites is dismissed as illusory in competitive equilibrium is based on what is now increasingly recognized as an outmoded Walrasian conception of contractual exchange. With the aid of new developments in the theory of moral hazard, principal-agent relations, transactions costs, and radical political economy, the Walrasian conception of exchange may be shown to be a limiting case of rather questionable importance. (Some of the relevant literature is cited in Joseph Stiglitz, 1987; Oliver Williamson, 1984; Gintis, 1976; and Bowles, 1987.)

James Buchanan has clearly identified a critical assumption of the underlying Walrasian model. Describing a prototypical exchange at his local roadside stand, Buchanan writes: "I do not know the fruit salesman personally, and I have no particular interest in his well-being. He reciprocates this attitude... Yet the two of us are able to complete an exchange expeditiously... We transact exchanges efficiently because both parties agree on the property rights relevant to them" (1975, p. 17).

When there is agreement between the parties to exchange concerning the relevant property rights, and when this agreement can be costlessly enforced, the neoclassical commitments follow. But, what if exchange must take place in the absence of (or under conditions precluding the enforceability of) such an agreement? In such cases we have a political problem. However, as Abba Lerner has observed, for neoclassical economics: "An economic transaction is a solved political problem. Economics has gained the title of queen of the social sciences by choosing solved political problems as its domain" (1972, p. 259).

Yet two of the exchanges most fundamental to the functioning of the capitalist economy have precisely the character of unsolved political problems. These are exchanges involving labor and financial capital. An employer offers a wage in exchange for which

the employee offers not some fully specifiable *pro quo*, but rather at best an unenforceable promise to perform at an adequate level of intensity, care, and initiative. The employer must develop a monitoring system to determine whether this level has been achieved in any particular case and must have means to discipline an employee whose performance is deemed unsatisfactory.

Similarly, a financial investor makes funds available to an enterprise, receiving *ex ante* neither a determinate return, nor even a specific probability distribution of returns, but rather the promise that the fiduciary obligations of the recipient will be dutifully carried out. The financial investor's interest is in this case protected only by instituting a costly array of rewards and constraints. The structure of financial intermediation, the legal regulation of securities markets, as well as customary economic practices specifying the conditions of creditworthiness, must be drawn upon to monitor the use of funds and to induce borrowers (such as the managers of firms) to act in the interest of financial investors.

Labor and finance capital are only the most important of a general category of goods that are subject to what we term *contested exchange*. An exchange is contested when some aspect of the good exchanged possesses an attribute that is valuable to the buyer, is costly to provide, and is at the same time difficult to measure or otherwise not subject to determinate contractual specification. In such cases, the *ex post* terms of exchange are determined by the monitoring, sanctioning, and incentive mechanisms instituted by the buyer to induce proper seller behavior. These systems for the endogenous enforcement of competing claims in the case of contested exchange are recognizable as elements of the economy's political structure.

In liberal societies, where the legitimate exercise of physical coercion is monopolized by the state, the instruments available to private economic agents for the endogenous enforcement of contracts are severely circumscribed. Among these instruments *contingent renewal* holds a central position: the buyer induces seller compliance by prom-

ising to renew the contract only if satisfied with the seller's performance. In Albert Hirschman's insightful terms, contingent renewal power is based on exit rather than on voice.

Nonrenewal of exchange is a threat, however, only if the buyer offers terms better than the seller's next best alternative. To render contingent renewal effective, then, the buyer must offer the seller what we term an *enforcement rent*. This rent persists in a competitive equilibrium for the simple reason that it results from the competitive optimizing behavior of the agents on the short side of the market. Those with the power to change the terms of the exchange have no incentive to do so; quantity constrained agents on the long-side of the market—the unemployed, the credit constrained—thus cannot compete away the rent.

For this reason, contested exchange markets do not generally clear in competitive equilibrium. Thus some agents have power over both price and quantity while others have power over neither. In general we may distinguish three types of agents: those on the short side of the market who use enforcement rents to as an instrument of control (for example, employers, financial intermediaries), those on the long side who succeed in making a transaction and thus receive the rents (the employed, the credit-worthy), and the long-siders who fail to make a transaction (the unemployed, the credit rationed). Because the short-siders are in a position to give commands which the long-siders are constrained to obey, and because some of the long-siders are quantity constrained (they cannot make a transaction), both markets and firms become arenas for the exercise of substantive economic power.

Not surprisingly, all three of the basic neoclassical propositions are false in the context of contested exchange.

## II. Property and Power: Enforcement as Market Failure

When Knut Wicksell demonstrated that in a constant returns to scale world, the results of a general competitive equilibrium would be unaltered by having agricultural workers

rent land, paying the landlords the marginal product of land and rewarding themselves with the residual rather than the converse, he assumed that the labor exchange had the usual Walrasian properties. Had he instead modeled the landlord-worker relationship as a contested exchange, it would have been perfectly clear that the location of the residual claimant status makes a very substantive difference. As has been widely recognized in the literature on hierarchical firm organization, owner-manager conflicts, and land tenure relationships, and as is confirmed by common sense, the enforcement problem arises because the agent controlling the contested variable (for example, work effort or managerial risk taking) is not the residual claimant and thus does not enjoy the full fruits and bear the full costs of his or her actions. But in the context of contested exchange, a reallocation of property rights generally involves a reallocation of residual claimant status. We thus affirm a counter-proposition, *the nonneutrality of property-assignments*: the reallocation of titles to property will generally alter the competitive equilibrium allocation, changing prices, enforcement costs, and output levels in contested exchange markets (and hence in all markets) as well as altering the distribution of income.

The nonneutrality proposition is true for yet another reason: the distribution of property rights alters the relative cost effectiveness of different endogenous enforcement mechanisms. If, for instance, workers own substantial amounts of property, they may post bonds as a condition of employment, thus reducing the costs of enforcing labor discipline. Again, if potential borrowers are wealthy, they may post collateral, thus reducing the monitoring and incentive costs of financial investors. In both cases, wealth redistributions alter competitive equilibrium allocations.

The neoclassical commitment to the irrelevance of the locus of command is also unwarranted. For in the case of contested exchange, the locus of command determines how contracts are endogenously enforced. Consider, for instance, a firm in contested exchange equilibrium in which the owner, has chosen a profit-maximizing configuration

of monitoring resources and enforcement-rents. Included in such a configuration is a profit-maximizing decision process involving an assignment of firm members to positions in the structure of command, an allocation of resources to monitor decision-making behavior, and a pattern of enforcement rents accruing to decision makers.

Now let us consider a new structure of command replicating the old in all respects save that additional firm members are permitted to participate at one or more points in the process of decision making. The owner might be able to reproduce the previously optimal decision by spreading monitoring resources across the additional decision makers, while increasing enforcement rents sufficiently to induce the original pattern of decisions. But, even assuming that the process of decision making per se does not expend real resources, this new locus of command entails an equilibrium transfer of wealth from the residual claimant to the newly empowered decision makers. The locus of command is thus not irrelevant, but has distributional effects even when the alternatives contemplated are limited to otherwise equivalent allocations.

The example can be significantly extended. Indeed, we may suggest a general *principle of induced autocracy*, that asserts that short-side agents in competitive financial markets will, under very general conditions, prefer to deal with firms controlled by a relatively small subgroup of members unaccountable to the general membership for its actions. Suppose the key operating parameters that a firm must choose have the following characteristics. First, these parameters render the relationship between the firm's decision makers and its financial investors one of contested exchange (i.e., the parameters are costly to monitor and decision makers are motivated by enforcement rents and contingent renewal). Second, the firm's performance depends upon the key parameters chosen, but not upon the process by which they were determined (i.e., the political process has no independent effect upon performance). Third, each decision maker's choice of key parameters is a function of his or her preferences and the incentives provided. Fourth, suppose the final val-

ues of the key parameters represent some convex combination (possibly degenerate) of the choices of members participating in the structure of command. Then financial investors will prefer a structure of command in which the number of participating decision makers is at most one greater than the number of key parameters to be chosen.<sup>1</sup> The irrelevance of command is thus false.

The third neoclassical proposition, the efficiency of market exchange, fares no better. Indeed, we may demonstrate the quite general *principle of Pareto-improving redistribution*: in any contested exchange equilibrium with positive monitoring costs, there will exist a redistribution of income entailing a Pareto-superior allocation of economic resources. This is the case because every cost-minimizing enforcement strategy will involve both enforcement rents—a distribution variable, and monitoring costs—a resource-using variable. Cost-minimizing claim enforcement is not efficient because it treats the resource-using and the non-resource-using enforcement instruments equivalently despite the fact that the former has social opportunity costs and the latter does not.

To see that this is the case, suppose that, starting from a cost-minimizing contested exchange equilibrium, the buyer increases the enforcement rent offered the seller by a small amount, while at the same time reducing the real resources devoted to monitoring, so as to maintain a constant level of seller compliance. Then the same quality of the contested good has been provided at a smaller expenditure of real resources despite the fact that (by construction) buyer costs have risen. Hence there must exist some redistribution between buyer and seller which renders both better off.

The principle of Pareto-improving redistributions contrasts sharply with the widely accepted “neo-Coasean” notion that competitive firms will choose “efficient” forms of monitoring. The flaw in the neo-Coasean

argument is simple: it fails to investigate explicitly the instruments available to the firm for disciplining wayward firm members. When these means involve contingent renewal, minimizing transactions costs does not entail minimizing transactions resources, as some instruments—the enforcement rents—while costly to the enforcer, bear no social opportunity costs. In the two cases we have focused upon—financial markets and labor markets—the implications are clear: by comparison to the profit-maximizing equilibrium, a small reduction in the interest rate or increase in the wage, followed by a suitable redistribution of property titles, would be Pareto improving.

### III. Economic Democracy and Property Distribution

Would-be contemporary constitution makers may well ponder these results. For they point to the importance of economic power, not only as a problem for democratic accountability, but as challenge to the putative economic rationality of competitive equilibrium in a private ownership economy. Three implications are particularly striking.

First, the argumentation surrounding the nonneutrality of property assignments suggests that enforcement costs will be minimized by locating residual claimant status in those agents supplying the service whose monitoring is relatively difficult. The fact that residual claimant status in the modern capitalist corporation resides in owners who generally supply little more than financial assets, the monitoring costs of which strikingly approach the Walrasian ideal, is simply irrational from the perspective of minimizing enforcement cost. A redistribution of ownership to employees of the firm might yield major efficiency gains.

Second, the principle of induced hierarchy demonstrates that worker-run and worker-owned firms will be shunned by financial markets despite a cost structure which on the above reasoning may be favorable by comparison with more hierarchical firms owned by the wealthy.

Third, while the transaction costs of democratic decision making may militate against economic democracy and a generalization of

<sup>1</sup>For instance, if there is one key parameter (for example, a point on a risk/expected return schedule), financial investors will prefer either a single decision maker, or two decision makers whose choices lie on opposite sides of the investors' optimum. See Gintis (1987) for a fuller statement and proof.

property ownership, these costs may be rather small compared to the high levels of enforcement cost in a capitalist economy, for as we have seen, these costs include not only by the direct monitoring of transactions in which the effective decision makers are rarely the residual claimants, but those associated with equilibrium credit rationing and unemployment as well.

The theory of contested exchange thus does more than illuminate the undemocratic political structure of the economy; it suggests that a competitive capitalist economy of hierarchically structured firms is not simply unequal, undemocratic, and inefficient, but rather that it may be inefficient *because* it is unequal and undemocratic.

Indeed, a stronger position might be sustained: that a more egalitarian and democratic economy would support greater acceptance of the rules of the game and less contestation over the distribution of economic reward, allowing a further reduction in enforcement costs, (this argument is made in Bowles, 1985). But to support this claim would take us beyond the concerns of this essay, to issues with which we have dealt elsewhere (1986).

#### REFERENCES

- Bowles, Samuel, "The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian, and Marxian Models," *American Economic Review*, March 1985, 75, 16-36.
- \_\_\_\_\_, "Contested Exchange: A Microanalysis of the Political Structure of the Capitalist Economy," mimeo., University of Massachusetts, 1987.
- \_\_\_\_\_, and Gintis, Herbert, *Democracy and Capitalism: Property, Community, and the Contradictions of Modern Social Thought*, New York: Basic Books, 1986.
- Buchanan, James, *The Limits of Liberty*, Chicago: University of Chicago Press, 1975.
- Gintis, Herbert, "The Nature of the Labor Exchange and the Theory of Capitalist Production," *Review of Radical Political Economics*, Summer 1976, 8, 36-54.
- \_\_\_\_\_, "The Principle of External Accountability in Financial Markets," Department of Economics, University of Massachusetts, November, 1987.
- Lerner, Abba, "The Economics and Politics of Consumer Sovereignty," *American Economic Review Proceedings*, May 1972, 62, 258-66.
- Stiglitz, Joseph E., "The Causes and Consequences of the Dependence of Quality on Price," *Journal of Economic Literature*, March 1987, 25, 1-48.
- Williamson, Oliver E., "The Economics of Governance: Framework and Implications," *Journal of Institutional and Theoretical Economics*, 1984, 140, 195-223.



## ISSUES IN THE BLACK COMMUNITY<sup>†</sup>

### Income, Wealth, and Investment Behavior in the Black Community

By ANDREW F. BRIMMER\*

Over time, as blacks' incomes have risen, they have increased their saving rate, and this has enabled them to enlarge their accumulation of assets. However, blacks' share of wealth is much smaller than their share of income.

The profile of asset accumulation by blacks generally parallels that in the economy at large. Yet, a few striking differences in asset preferences—distinguished by the degree of risk involved—are evident when blacks' portfolios are compared with those held by whites.

Most blacks have little knowledge of the stock market, and they have developed only a weak demand for equity securities. However, over time, as their incomes rise further and as they acquire more familiarity with common stocks, more blacks will include these issues in their investment portfolios.

These general conclusions are amplified in this paper. In Section I, the level and composition of assets owned by blacks are compared with the pattern for whites and the country at large. Particular attention is given to the distribution of assets as a mirror of blacks' attitudes toward risk. The role of black investors in the stock market is discussed in Section II.

#### I. Income, Wealth, and Asset Choices

The accumulation of wealth by blacks falls substantially short of what one might have

expected from a familiarity with the money income figures available annually. The size of the deficit in black wealth can be measured on the basis of data from the U.S. Bureau of the Census (see U.S. Department of Commerce, 1986).

Using these figures, detailed estimates of the amount and composition of wealth held by households in the United States in 1984 were prepared. The estimates are shown in Table 1. In compiling the data, the Census Bureau did not include the assets that individuals have accumulated in the form of pension funds, life insurance policy reserves, and Social Security contributions. The Census Bureau data show "net worth," which represents the market value of the specified assets owned by households minus the households' total liabilities.

#### A. Level of Accumulated Wealth

In 1984, the wealth of the black community amounted to \$208.2 billion. Thus, blacks owned 3.0 percent of the \$6,912.2 billion of accumulated wealth in the United States. By comparison, in 1984, blacks received 7.2 percent of the nation's total money income. Therefore, their share of wealth was less than half their share of income.

In 1984, there were 9,509,000 black households in the country, equal to 11.0 percent of the total of 86,790,000. In that year, black money income amounted to \$171.6 billion—7.2 percent of the total. If blacks had received their proportionate share of income (11.0 percent), they would have gotten \$263.1 billion—or \$91.4 billion more than they actually received. These figures imply an income deficit of 34.8 percent. Their deficit in wealth was much larger. If blacks had owned 11.0 percent of the total wealth in 1984, their

<sup>†</sup>*Discussants:* Marcus Alexis, University of Illinois-Chicago; Margaret Simms, Joint Center for Political Studies; William Darity, Jr., University of North Carolina-Chapel Hill.

\*President, Brimmer & Company, Inc., and Wilmer D. Barrett Professor of Economics, University of Massachusetts, Amherst, MA 01003.

TABLE 1—MONEY INCOME AND NET WORTH,  
BY TYPES OF ASSETS OWNED,  
TOTAL AND BLACK HOUSEHOLDS, 1984

Asset Type	Total (1)	Black Households	
		(2)	(3)
Reg. check accounts	43,131	1,823	4.2
Interest-earning assets at fin. insts.	985,512	13,057	1.3
IRA/Keogh accounts	149,794	1,669	1.1
U.S. savings bonds	32,443	387	1.2
Other interest-earning assets	213,653	1,735 <sup>a</sup>	0.8
Stocks and mutual fund shares	466,357	1,443	0.3
Equity in business or profession	706,501	12,919	1.8
Motor vehicles	410,754	21,300	5.2
Homeownership	2,822,332	124,592	4.4
Vacation homes and other real estate	298,853	4,529	1.5
Rental property	613,389	23,953	3.9
Other assets	169,484	782 <sup>a</sup>	0.5
<i>Total: Net Worth</i>	6,912,202	208,189	3.0
Memorandum:			
Money Income	2,391,693	171,649	7.2

Source: Analysis and calculations by Brimmer & Company, Inc. Basic data from U.S. Department of Commerce, Bureau of the Census, Tables 1, 2, and 3.

Note: Cols. 1 and 2 are amounts, shown in millions of dollars; col. 3 is percent of the total.

<sup>a</sup>Estimated by Brimmer & Company, Inc.

net worth would have amounted to \$760.3 billion—or \$552.2 billion more than they actually held. Thus, their wealth deficit amounted to 72.6 percent.

### B. Types of Assets

The assets in Table 1 are arrayed roughly by liquidity. The first four categories represent primarily bank accounts and liquid savings. In combination, they accounted for 17.7 percent of the net worth of all households in the country in 1984. The corresponding fraction was 8.2 percent for black households. The second group consists of financial investments. The interest-earning assets in this group include mainly money market funds, U.S. government securities, municipal obligations, and corporate bonds. Stocks and mutual fund shares represent household ownership of corporate enterprises. Blacks'

investment in these categories was quite meager. They held only 0.8 percent of the bonds and money market funds and only 0.3 percent of stocks and mutual fund shares. Such assets also accounted for a small fraction (0.8 percent) of blacks' total wealth.

In a similar vein, the equity accumulated by blacks in the form of business or professional assets was quite modest. Their share amounted to 1.8 percent of the total held by all households. This category represented 6.7 percent of the total wealth of black households—a proportion only three-fifths of the 10.3 percent recorded for all households in the country.

The ownership of physical assets is reflected in the value of motor vehicles, homes, vacation homes, rental property, and other real estate. In 1984, blacks owned 5.2 percent of the motor vehicles held by households—which represented their largest share of all of the different types of assets shown in Table 1. Motor vehicles also accounted for 11.1 percent of blacks' total wealth. The corresponding fraction was 6.0 percent for all households.

As one would expect, the equity accumulated in their homes represented the most important form of wealth held by blacks. This equity was valued at \$124.6 billion in 1984 and equaled 4.4 percent of the value of all homes in the nation. For blacks, homeownership accounted for 65 percent of their total wealth compared with 40 percent for all households in the country. Blacks owned only 1.5 percent of the vacation homes and other real estate. The value of these properties represented 2.2 percent of blacks' wealth vs. 4.4 percent of the wealth of all households combined. Rental property is relatively more important as a form of investment for blacks than it is for all households in the country. Blacks owned 3.9 percent of the rental property, and the latter accounted for 12.4 percent of their total wealth. For all households the figure was 8.6 percent.

### C. Income and Wealth

The data in Table 2 show the distribution of wealth by level of income. In 1984, the median annual income of all households was \$20,124. The median income for whites was

TABLE 2—MEDIAN NET WORTH BY RACE  
AND ANNUAL HOUSEHOLD INCOME, 1984

Category	Households		
	All	White	Black
Median Income	\$20,124	\$21,120	\$13,056
Median Net Worth	32,667	39,135	3,397
Annual Income			
under \$10,800	5,080	8,443	88
\$10,800–23,999	24,647	30,714	4,218
\$24,000–47,999	46,744	50,529	15,977
\$48,000 and over	123,474	128,237	58,758
Type of Household			
Married Couples	50,116	54,184	13,061
Female Householders	13,885	22,500	671
Male Householders	9,883	11,826	3,022

Source: Calculations by Brimmer & Company, Inc. Data from U.S. Department of Commerce, Bureau of the Census, Tables 1, 2, and 3.

\$21,120, and that for blacks was \$13,056. Thus, black income was 61.8 percent of that for whites. Also in 1984, the median net worth of all households was \$32,666. The corresponding figures for white and black households were \$39,135 and \$3,397, respectively. Therefore, median black wealth was 10.4 percent of the amount owned by whites.

As one would expect, the relative position of blacks with respect to wealth accumulation varied with income. Among households with annual incomes under \$10,800, the median net worth of white households was \$8,443. The corresponding figure for black households in the same income class was \$88—only 1.0 percent of the white median amount. In the income range of \$10,800 to \$23,999, the median wealth of white households was \$30,714 compared with \$4,218 for black households, resulting in a black-white proportion of 13.7 percent. White households with median incomes between \$24,000 and \$47,999 had a median net worth of \$50,529. The parallel figure for blacks was \$15,977—yielding a black-white percentage of 31.6. At the upper end of the income scale (in excess of \$48,000), the median net worth of white families amounted to \$128,237, and that for black families amounted to \$58,758. This represented a black-white wealth proportion of 45.8 percent.

The distribution of wealth was even more striking when the asset holdings of different

types of households were compared. For instance, in 1984, white married couple households had a median net worth of \$54,184. Among black married couple households, the median net worth was \$13,061. The black-white fraction was 24.1 percent. In sharp contrast, white female householders had a median net worth of \$22,500, and the corresponding figure for black female householders was \$671. In this case, the black-white proportion was 3.0 percent. Among male householders, the median net worth for whites was \$11,826, and it was \$3,022 for blacks. These figures yielded a black-white wealth relation of 25.6 percent.

The distribution of wealth within the black community is far more uneven than it is among whites in the nation at large. For example, among all households combined, the net worth of those at the top of the income scale was 24 times as large as the net worth of those in the lowest income category. Among white households, the multiple was 15 times, but among blacks it was 67 times. A similar pattern—though less extreme—prevailed when types of households are compared. Thus, the wealth held by married couples in the nation at large was nearly 4 times that held by female householders. Among whites, married householders had about 2.5 times the wealth held by female householders. In sharp contrast, the wealth held by black married couples was 19.5 times as large as that held by black female householders.

In summary, the ownership of wealth by blacks reflects the same pattern of deficits evident when one looks at money income. However, the shortfall in wealth is much larger. To a considerable extent the latter can be traced to a long history of deprivation in this country. This means that blacks have had much less opportunity than whites to earn, save, or to inherit wealth. Because of this historical legacy, black families have had few opportunities to accumulate wealth and to pass it on to their descendants.

## II. Blacks in the Stock Market

Historically, the stock market has attracted very few black investors. However, a sprinkling of black individuals has appeared

in the market, and the number appears to be growing somewhat.

#### A. Stock Ownership by Blacks

The first comprehensive estimate of stock ownership by blacks was obtained from the *Survey of Economic Opportunity (SEO)* conducted in 1967 by the Office of Economic Opportunity (OEO) in the Johnson Administration. The canvass estimated that stocks held by blacks were valued at \$200 million (see Henry Terrell, 1970). At the same time, all families in the country taken together owned stocks valued at \$145.2 billion. These figures suggest that blacks owned 0.13 percent of the total amount of stock reported. In the same year, the total net wealth accumulated by blacks was estimated at \$22.7 billion. This figure represented 2.0 percent of the \$1,134.7 billion of net wealth reported by all families. Expressed differently, corporate stocks accounted for only 0.9 percent of blacks' net wealth holdings. The corresponding figure for all families was 13.1 percent.

To obtain a more recent profile of stock ownership by blacks, Brimmer & Company made an estimate of the amounts held at the end of 1982 (see my 1983 paper). The calculations used as benchmarks the statistical relationships between income and wealth accumulation by race derived from the 1967 *SEO* data. These coefficients measure the percentage change in net wealth associated with a 1.0 percentage change in money income. Data on the ownership of corporate stocks and other financial assets by all households at the end of 1982 were obtained from the Federal Reserve Board. The income-wealth coefficients for blacks derived from the *SEO* data were revised to allow for the fact that blacks' income rose faster (by 1.14 times) than that for the population at large.

The results of these calculations can be summarized. At the end of 1982, the total wealth of the black community was estimated at \$144.6 billion, and the amount for the country at large was \$6,189.7 billion. Thus, blacks held 2.34 percent of the total. Blacks had money income of \$148.0 billion in 1982. In the same year, total money in-

come was estimated at \$2,055.8 billion, of which blacks got 7.2 percent.

The value of corporate equities owned by all households at the end of 1982 was reported at \$1,310.8 billion. The amount of stocks owned by blacks at the end of that year was estimated by Brimmer & Company at \$3.3 billion. Thus, blacks held 0.25 percent of the amount owned by all households. These calculations also indicate that corporate stocks represented 2.3 percent of the total wealth held by blacks, and 21.1 percent of that held by all households.

#### B. Profile of Black Investors

The reasons why the stock market has traditionally not attracted many blacks—as well as the factors which are leading to somewhat increased participation by them—stand out clearly in the results of informal surveys of stock brokers and investment counselors conducted by Brimmer & Company. The original canvass was undertaken in 1983 and updated in the fall of 1987.

In general, the fairly low average income in the black community means that the typical black family has had a low saving rate. However, in recent years, the percentage of disposable income saved by blacks has been rising, and more of the extra cash is being invested.

Investors' attitudes toward common stock ownership are shaped by a variety of factors—including the degree of familiarity with securities markets and the perception of risk. For the most part, blacks know very little about the stock market and the opportunities for capital accumulation which the latter offers. Much of that type of information is acquired through business and social contacts with stock brokers and others active in the financial arena. Very few blacks have such contacts. This same relative isolation makes it difficult for blacks to appraise the risk of loss that is inherent in the ownership of common stocks. As a result, blacks who do have a margin of funds to invest typically prefer to acquire much safer assets—such as savings accounts or real estate.

The personal characteristics of black common stock investors do not differ substan-

tially from those of their white counterparts. For example, their age range is quite similar—with active buyers in both groups being predominantly in their late 30s to late 40s. On the other hand, when the total ownership of stocks is distributed by age and sex, it is clear that older white women own a disproportionate share of the amount outstanding. This pattern is primarily a reflection of the fact that many of them inherited their holdings from late husbands.

Most blacks who are currently purchasing common stocks have incomes in the range of \$40,000–\$50,000. Among whites, the range is quite similar. However, proportionately more whites than blacks with incomes in the next bracket down (\$30,000–\$40,000) are active in the stock market. To some extent, these investors probably receive assistance from their parents. The stock brokers covered by Brimmer & Company's informal canvass indicated that they had very few black investors who had either very high income (in excess of \$100,000) or large accumulations of wealth (in excess of \$500,000). As a result, they have very few black clients who are substantial investors in common stock. When they do enter the market, blacks seem to concentrate relatively more on securities issued by companies with the highest credit ratings.

Brokers reported that, among black investors, young women are much more active in the stock market than are black men in the same age and income categories. The brokers had no ready explanation of this differential pattern of participation. But several did suggest that many single black women (in their late 30s) with steady jobs and good incomes are likely to have a larger margin of savings than would be true of their male counterparts. They also seem much more willing to make the effort necessary to become familiar with the stock market. They appear somewhat more venturesome and more ready to take risk—although their attitude toward the

latter remains rather cautious. In combination, these factors have induced relatively more black women than black men to become common stock investors.

Black common stock investors are still drawn substantially from among persons engaged in professional occupations (especially from among physicians, lawyers, corporate officials, and higher-paid government workers). Very few black entrepreneurs are substantial investors in the stock of publicly held corporations. Instead, they appear to prefer plowing back whatever profits they have in their own enterprises.

Looking ahead, one can expect blacks to become somewhat more active buyers of common stock. Some of these purchases will be made through brokers—among whom a few more blacks are appearing each year. Blacks will also acquire more shares through mutual funds, company-sponsored savings plans, payroll deductions, and the establishment of individual retirement accounts. This increased participation in the market will strengthen black investors' self-confidence—which, in turn, will increase their demand for common stock.

## REFERENCES

- Brimmer, Andrew F., "Blacks in the Stock Market," *Black Enterprise*, October 1983, p. 41.
- , "Investing and Wealth Accumulation," *Black Enterprise*, October 1986, p. 37.
- Terrell, Henry S., "Wealth Accumulation of Black and White Families: The Empirical Evidence," paper presented before Joint Session of the AEA and AFA, Detroit, Michigan, December 28, 1970.
- U.S. Department of Commerce, Bureau of the Census, "Household Wealth and Asset Ownership: 1984," *Current Population Reports*, Household Economic Studies, Series P-70, No. 7, July, 1986.

# The Social Preference for Fair Housing: During the Civil Rights Movement and Since

By WILHELMINA A. LEIGH\*

This paper applies the theory of social preference to interpret the results of opinion polls that included questions on the equality of access to housing, otherwise known as fair housing. Using social preference theory, a set that is decisive on the issue of fair housing for the years of the civil rights movement—designated as 1954 through 1968—and for the years since then is defined. Results of opinion polls with responses from whites to pro-integrative residential moves by blacks are used to estimate simple linear regression equations. These equations are, in turn, used to project the date when 90 percent of the set that is decisive on fair housing would favor these pro-integrative moves.

The motivation for this analysis was to test the assessment by strategists of the civil rights movement that fair housing was one of the goals least likely to be achieved by relying on the methods that had been effective to achieve other goals of this movement. In specific, the eradication of involuntary neighborhood segregation was deemed unlikely because it was not vulnerable to two of the major tools of the movement—court decisions and nonviolent direct action. (See M. Viorst, 1979, for more discussion of the major strategists of the civil rights movement and their tactics.)

The paper is organized as follows. Section I provides the basic theorems of preference theory for the individual and for society. Section II discusses the decisive set for the issue of fair housing both during the civil rights movement and since. Sections III and IV interpret opinion polls using preference theory, and the final section synthesizes the analysis.

\*Principal Analyst, U.S. Congressional Budget Office (CBO), Washington, D.C. 20515. The analysis and conclusions in this paper are mine alone. They should not be attributed to the U.S. CBO.

## I. Preference Theory

The theory of individual revealed preference is based on the following: 1) If the behavior of a consumer expressing his or her preferences over all binary combinations (or his relation over choices between paired alternatives) conforms to certain simple axioms, then the existence and nature of his indifference map or utility function can be inferred from his actions. 2) This binary relationship of preferences for all the pairs of alternatives is an ordering.

A preference ordering (or binary relationship that is an ordering) can be expressed either by  $a_1Ra_2$ , meaning " $a_1$  is at least as well liked as  $a_2$ ," or by  $a_1Pa_2$ , meaning " $a_1$  is preferred to  $a_2$ ." The ordering  $P$ , a derivative of  $R$ , is often called a strong preference ordering, while  $R$  is labeled a weak one. The axioms that the preference ordering must satisfy for all binary combinations in the set of alternatives,  $A$ , include: completeness, reflexivity, transitivity, and connectedness. (See J. M. Henderson and R. E. Quandt, 1971, for a thorough discussion of these axioms.)

Social preference theory establishes conditions under which a group decision procedure—known also as a collective choice rule or a collective choice function—can be defined, given certain data on the preference (or utility) functions of the individual consumers who comprise the group. A collective choice rule is a mapping from the series of choice rules  $R_1, R_2, \dots, R_n$  for the  $n$  individuals or consumers to a single rule,  $R$ , for the individuals collectively. If  $R$  is an ordering defined over a finite set  $A$  (of alternatives), then a choice function  $C(S, R)$  is defined over  $A$  for every nonempty subset  $S$  of  $A$ . (Discussion follows Amartya Sen, 1970.)

Kenneth Arrow asked the following question—given the preference orderings of  $n$  consumers, is there some collective choice

rule or aggregation procedure (known as a group preference function or ordering) by which society reaches its decisions? Arrow answered this question with his Impossibility Theorem—it is not possible to derive a general preference ordering,  $R$ , from a group decision procedure that satisfies the following four conditions: domain, unanimity, nondictatorship, and the independence of irrelevant alternatives.<sup>1</sup>

Establishing the conditions to derive a collective choice rule for all consumers is the main contribution of Sen to the social choice literature. The definition of quasi transitivity for the relation  $R$  allows this derivation. A relation  $R$  is transitive if the following maintains for all commodity combinations: if  $a_1Ra_2$  and  $a_2Ra_3$ , then  $a_1Ra_3$ . Quasi transitivity exists when  $R$  is a quasi ordering and meets certain conditions less stringent than for  $R$  as an ordering.<sup>2</sup> The quasi-ordering  $R$  is quasi transitive if for all  $a_1, a_2, a_3$  in the set  $A$ ,  $a_1Pa_2$  and  $a_2Pa_3$  yields  $a_1Pa_3$ .

Sen modified social preference theory to allow the definition of a collective choice rule for society, but this rule cannot be associated with a set of social indifference curves. In other words, a collective choice rule exists if the binary relation  $R$  is defined as a quasi ordering rather than an ordering.

Sen also defines a particular kind of collective choice rule, the social decision function (SDF). He proves that if  $R$  is reflexive, complete, and quasi transitive over a finite

set  $A$ , then a choice function  $C(S, R)$  is defined over  $A$ . Sen then defines a SDF as a collective choice rule  $f$ , the range of which is restricted to those preference relations  $R$ , each of which generates a choice function  $C(S, R)$  over the whole set of alternatives  $A$ . He goes on to demonstrate that there is a SDF that satisfies Arrow's four conditions—domain, unanimity, nondictatorship, and the independence of irrelevant alternatives. Sen notes, though, that the social preference relation  $R$  generated by the SDF and that satisfies Arrow's four conditions is merely quasi transitive.

The final social preference concept of interest is the decisive set. A set of individuals, " $O$ ", is decisive for  $a_1$  against  $a_2$  if  $a_1Pa_2$  when  $a_1Pi a_2$  holds for every individual,  $i$ , in " $O$ ". Note that there can be more than one decisive set of consumers in a society and that all decisive sets need not be decisive on all issues.

## II. The Decisive Set for Fair Housing

To examine the issue of fair housing, let us consider a society with  $n$  consumers or individuals, grouped as follows: 1)  $1, \dots, b$  are black. This group includes residents of all types of neighborhoods—open, moderately integrated, substantially integrated, integrated but in localities with few blacks, integrated in rural areas, substantially segregated, and completely segregated neighborhoods. (See N. Bradburn, S. Sudman, and G. Gockel, 1970, for a discussion of these neighborhood classifications.) 2)  $b+1, \dots, w$  are white and can be found living in neighborhoods defined similarly to those occupied by blacks, and 3)  $w+1, \dots, n$  are members of all other racial groups and can live in the same types of neighborhoods as blacks and whites.

Note that these sets are defined only along racial lines, and the many different subsets that would result if the groups were further categorized by income or employment are not considered. When discussing decisive sets below, however, subsets of these racial groups that are employed by the various branches of government (legislative, executive, and judicial) will be considered, as will subgroups

<sup>1</sup>The domain or completeness condition means that an aggregation procedure is defined for all  $n$ -tuples of individual orderings or mappings ( $R_1, R_2, \dots, R_n$ ). The unanimity condition means that there exists  $(a_1, a_2)$  in the set  $A$  such that for all  $i$ ,  $(a_1, a_2)$  in the set  $T^i$  implies that  $(a_1, a_2)$  is in the set  $T$ . Here the set  $T$  refers to those preferred choices from the set of available elements,  $A$ . The nondictatorship condition establishes that for any  $i$ , there is an  $(a_1, a_2)$  in the set  $A$  such that  $(a_1, a_2)$  in  $T^i$  and  $(a_2, a_1)$  not in  $T^i$ , and, for all  $j$ ,  $(a_2, a_1)$  in  $T^j$  and  $(a_1, a_2)$  not in  $T^j$  imply that  $(a_2, a_1)$  is in  $T$ . The independence of irrelevant alternatives means one can neglect any differences due to changed tastes when the alternatives are irrelevant.

<sup>2</sup>Conditions for a quasi ordering to exist are  $a_1Ia_2$  and  $a_2Ia_3$  yields  $a_1Ia_3$ ;  $a_1Pa_2$  and  $a_2Ia_3$  yields  $a_1Pa_3$ ;  $a_1Ia_2$  and  $a_2Pa_3$  yields  $a_1Pa_3$ ; and  $a_1Pa_2$  and  $a_2Pa_3$  yields  $a_1Pa_3$ , where  $I$  reflects indifference and  $P$  reflects preference. See Lemma 1\*a in Sen (p. 10).

employed in the real estate sector. Note also that the set of individuals  $w + 1, \dots, n$  of all other racial groups was a small set before the civil rights movement. Since the civil rights movement, its size has increased noticeably, and some of its members (mainly Hispanics) are encountering impediments similar to those faced by blacks seeking equal opportunity in access to housing.

Each member of each of these groups has a preference ordering,  $R_i$ , over any set of alternatives,  $A$ , that is governed by the theory of individual preference. The individual preference orderings (i.e., the  $R_1, R_2, \dots, R_n$ ) can yield a collective choice rule for each subgroup of individuals, under certain conditions. If the relation  $R$  for each of these sets of individuals collectively is quasi transitive, then an associated collective choice rule, such as the SDF, is guaranteed to exist, although a set of indifference curves for society cannot be associated with this function.

The history of fair housing in the United States suggests that the set of consumers that has been decisive has been dominated by members from the group of white individuals,  $b + 1, \dots, w$ , who have functioned to limit equality of access to housing for blacks both as private citizens and in their occupational roles. Although the poll results reported in the next sections substantiate that not all whites have preferred to proscribe the access of blacks to housing, the subset of whites that has been decisive on this issue has acted to do so. The decisive set on fair housing has been dominated by members from the set  $b + 1, \dots, w$  (for example, the subset  $b + 1, \dots, w - d$ , where  $d$  is a small number) and these members have preferred not to establish equal access to housing for all racial groups in society.

Prior to the years of the civil rights movement, the decisive set for the SDF on fair housing was comprised of the white members of the real estate profession, the whites in decision-making roles in all the branches of federal, state, and local government (i.e., legislative, executive, and judicial), and many white citizens. They conspired to prevent equal access to housing for blacks and were

successful with the assistance of such tools as the doctrine of "separate but equal," racial zoning ordinances, racially restrictive covenants on deeds, and a series of court decisions that allowed agents of the government to continue ignoring statutes intended to guarantee rights to blacks.

During the civil rights movement, the membership of society's decisive set on the issue of fair housing seems to have been reduced. A number of actions were taken at the federal level of government intended to guarantee equality of access to housing. In 1962, Executive Order 11063 was issued to outlaw racial discrimination in certain types of federally assisted housing. The legislative branch of the federal government passed the Civil Rights Act of 1964 and the Civil Rights Act of 1968, the latter with its Title VIII designated the Fair Housing Act. In a 1968 decision, the Supreme Court affirmed the Civil Rights Act of 1866 that precluded discrimination by race in the sale or rental of real property. Thus, the decisive set against fair housing consisted primarily of some individual whites, members of the state and local governing bodies toward which the civil rights protests were targeted, and the predominantly white real estate profession.

Since the end of the civil rights movement, although the emphasis on fair housing as a goal has increased, the membership of the decisive set against fair housing has increased. At the same time that the Supreme Court made decisions in favor of residential integration, the executive branch failed to issue interpretive regulations for Title VIII of the 1968 Act. The legislative branch of the federal government has failed in its attempts to amend and strengthen Title VIII, while testers have revealed continued housing market discrimination. (See the article by H. Newburger, 1984, and the report by R. Wienk et al., 1979, for a discussion of the use of testers and their findings from selected cities.) In the 1970's and 1980's, the decisive set against fair housing continues to consist of some white individuals, the mainly white decision makers in the state and local governing bodies, and members of the real estate profession. In addition, the decisive



set seems now to contain the power players of both the legislative and executive branches of federal government.

Although black households have not been decisive on the issue of fair housing, their preferences can be less clearly discerned from their behavior than the preferences of whites can be. Fair housing may or may not imply residence in integrated neighborhoods—that is, neighborhoods that both blacks and whites can move into and that both groups are continuing to move into. The term could encompass residence in integrated neighborhoods and also imply the access to market neighborhoods independent of their racial composition. Discrimination thus would be defined as behavior that denied to any group the right or opportunities given to others when seeking a neighborhood into which to move.

Blacks have lived in segregated neighborhoods throughout the United States for many decades. (See the work by A. B. Schnare, 1980, and by A. Sorensen et al., 1975, both of which measure the extent of this segregation.) Residence in segregated neighborhoods may reflect equal opportunity in access to housing, however, if people have voluntarily segregated themselves by race. Because so many tactics have been used to involuntarily segregate blacks, it is difficult to measure the degree of actual black self-segregation.<sup>3</sup> Segregation becomes the opposite of fair housing when blacks are forced to reside in segregated neighborhoods because of governmental actions and market machinery that exclude them from other areas containing housing for which their economic circumstances would allow them to compete.

### III. Interpreting Opinion Polls During the Civil Rights Movement

The poll results cited in this and the following section were tabulated by the major national survey research organizations—

Gallup, Harris, the Institute for Survey Research (ISR), the National Opinion Research Center (NORC), and Roper. (See the works by J. M. Goering, 1986, and M. A. Schwartz, 1967, for the survey results used in this analysis.) The samples of respondents and the phrasing of questions differ somewhat among surveys over the years, although the concepts underlying the questions are unchanged. Results are reported for the most clearly defined response category—usually the negative one—for questions generally phrased to elicit if one would be bothered by certain neighborhood racial residential patterns.

One thing to bear in mind with these poll results is that expressed opinions are often more liberal or favorable toward equal access to housing for blacks than are actions taken. Frequently, a decisive set determined on the basis of action taken would differ from a decisive set defined on the basis of poll results.

During the civil rights movement, pollsters surveyed whites on their attitudes toward race and housing and their perceptions of the major objectives of the civil rights movement. Respondents were queried about their subsequent behavior if blacks were to move next door, to move in great numbers to their neighborhoods, and to move onto their blocks.

In 1958, 1963, 1965, 1966, and 1967, Gallup asked its respondents whether they would move if black people came to live next door. The percentage of whites who replied "No" went from 56 in 1958 to 63 in 1967 (with a slight decrease to 55 percent in 1963). A straight line was fitted, with the percentages of whites who would not move as a function of the survey years, 1958–67. The resulting equation is

$$(1) \quad Y = 1.06X - 2011.44 \quad R^2 = .61. \\ (2.15)$$

(where the number in parentheses in this and all subsequent equations is the *t*-statistic).

The decisive set that would be operative if fair housing legislation were enforced so that housing market racial discrimination were

<sup>3</sup>For evidence that black residential segregation is not a reflection of black preferences, see Joe Darden (1983, p. 22).

eliminated is here defined as "those whites who would not move if blacks moved next door." When 90 percent of the survey respondents indicated they would not move under these circumstances, society would be close to ridding itself of housing market racial discrimination. Using the estimated equation to project to the future, the year in which 90 percent of respondents would indicate they would not move was determined. The year was 1992, a result that seems plausible because of the downturn in segregation between 1970 and 1980 (reported both by S. McKinney and Schnare, 1986, and by K. Taeuber, 1983). The year 1992 is only 4 years away, however, and it is difficult to imagine that the impact of increased membership in the decisive set "against" fair housing during the 1980's would be to achieve a decisive set that is "for" fair housing so soon.

When asked (also by Gallup and in 1958, 1963, 1965, 1966, and 1967) whether they would move if blacks came in great numbers to live in their neighborhoods, the percentage of negative responses rose by 8 percentage points over that period, from 20 percent to 28 percent. The high point for this period occurred in 1966, when 30 percent said they would not move. Using the approach indicated above, the following equation showed that it would be the year 2023 before the desired decisive set of whites who indicated they would not move from a neighborhood if great numbers of blacks moved in would contain 90 percent of all survey respondents. The equation estimated was

$$(2) \quad Y = 1.08X - 2092.62 \quad R^2 = .86. \\ (4.28)$$

It is hard to imagine what race relations would be like 37 years from now, but the fact that the projected date is so distant tells us that whites are reluctant to live in neighborhoods in which they would be in the minority.

The "same block" question asked in 1942 by NORC was repeated in 1956, 1963, 1964, 1965, 1966, and 1968. The percentage of white respondents who would not move if

blacks with the same incomes and education moved onto their blocks, increased dramatically over those years from 53 percent in 1956 to 77 percent in 1968. The linear regression equation based on this poll data is

$$(3) \quad Y = 1.49X - 2850.86 \quad R^2 = .96. \\ (11.54)$$

The date when a decisive set of 90 percent of white respondents would not move if blacks with equivalent incomes and education moved onto their blocks, turned out to be 1980. Although I have not examined the income and educational compatibility of black and white neighbors on the same blocks in 1980, the fact that these data project to a data that has already passed, suggests that the most acceptable form of racial integration is of a black family perceived to be equal in education and income on one's block—but not next door.

#### IV. Interpreting Opinion Polls Since the Civil Rights Movement

Opinion polls taken since the civil rights movement have included fewer questions on race relations in general and on racial attitudes toward housing in particular. The last year in which the NORC asked the "same block" question was 1972, when 85 percent of white respondents indicated they would not move if a black with the same income and education moved onto their block. The linear regression equation for this question estimated for the years 1942 through 1972 is

$$(4) \quad Y = 1.59X - 3050.98 \quad R^2 = .96. \\ (11.63)$$

In 1978, the percentage of households who would not move was 90 percent, again confirming the relative preferability among whites of this model for residential racial integration.

The last year in which the Gallup poll asked respondents about their behavior if blacks were to move next door and if blacks were to move into their neighborhoods in

great numbers was 1978. At that time, 86 percent said they would not move if blacks moved next door, and 46 percent said they would not move if blacks moved into their neighborhoods in great numbers. The linear regression equations estimated for these questions are, respectively,

$$(5) \quad Y = 1.62X - 3117.27 \quad R^2 = .91; \\ (6.31)$$

$$(6) \quad Y = 1.34X - 2610.85 \quad R^2 = .97. \\ (10.51)$$

In 1982, 90 percent of whites should have been willing to remain if blacks had moved next door to them. Only if there have been no instances between 1982 and the present, and if there turns out to be no future evidence of whites leaving a neighborhood when a black family moves next door, can we believe that the decisive set was changed as predicted in 1982. If the figure were really 90 percent in 1982, it should certainly be so close to 100 percent by now that white flight would no longer be measurable. It would be the year 2011 before 90 percent of whites would remain if great numbers of blacks moved into their neighborhoods.

Because of the poll results examined and the definition of a decisive set, one can see that it will be a long time before black consumers  $(1, \dots, b)$  are members of the decisive set and before the decisive set favors fair housing. In the period since the civil rights movement, the degree of residential segregation has declined nationwide. However, when not quite half of all white respondents favored open housing laws as recently as 1983, I still see a decisive set of white consumers  $(b + 1, \dots, w - d)$  that could lessen the access to housing opportunities for blacks by influencing the implementation of the state and federal fair housing statutes now on the books. Once 100 percent of whites answer "No" when polled on questions such as those presented above, then there may be some possibility for blacks  $(1, \dots, b)$  and other minority racial groups  $(w + 1, \dots, n)$  to become part of the decisive set.

## V. Synthesis of Analysis

Despite some changes in attitudes about blacks as neighbors, the preferences of society during the civil rights movement were revealed by a decisive set consisting of many white consumers, many legislators, and many in the executive and judicial branches of government. Although executive orders and regulatory changes emanated from the executive branch and legislators passed a series of civil rights laws during the period, the sentiment of the decisive set seemed to be to "make haste slowly" to dismantle the "separate but equal" society. One observer has noted that of the two decades the civil rights movement straddled, some part of any changes in the residential patterns of blacks between 1960 and 1970 may be attributable to executive orders and legislation related to fair housing, while none of the changes observed between 1950 and 1960 can (see D. Lockard, 1968).

In the years since the civil rights movement, the use of testers has revealed continued maneuvering by real estate professionals to exclude and segregate blacks and other minorities. Although income disparities continue to limit the degree to which equality in housing status can be attained, they are not responsible for the extent of residential racial segregation in our society (see U.S. Commission on Civil Rights, 1983). In spite of the scarcity of poll results from the years following the civil rights movement, the decisive set that can be identified as responsible for maintaining the disparate housing status quo in society consists of various subsets of white consumers—real estate professionals, the members of the legislative branch who seek to subvert or dilute the intent of existing civil rights statutes, the members of the executive branch who are unwilling to operate fair housing enforcement programs as mandated, and the members of the judicial branch who hand down decisions inimical to increased racial equality in access to housing.

The decisive set may be losing some members among private citizens, however, since in 1978, 86 percent of whites polled said they would not move if blacks moved next door.

On the other hand, only 46 percent of the whites polled in 1983 indicated that they favored open housing laws. This lack of popular support for existing legislation may be what encourages the individuals in the executive and legislative branches to try to subvert the intent of existing statutes.

## REFERENCES

- Bradburn, N., Sudman, S. and Gockel, G., *Racial Integration in American Neighborhoods*, Chicago: National Opinion Research Center, 1970.
- Darden, J., "Population Growth and Spatial Distribution," in *A Sheltered Crisis...*, U.S. Commission on Civil Rights, 1983.
- Goering, J. M., *Housing Desegregation and Federal Policy*, Chapel Hill: University of North Carolina Press, 1986, ch. 7.
- Henderson, J. M. and Quandt, R. E., *Microeconomic Theory: A Mathematical Approach*, New York: McGraw-Hill, 1971.
- Lockard, D., *Toward Equal Opportunity: A Study of State and Local Antidiscrimination Laws*, New York: Macmillan, 1968.
- McKinney, S. and Schnare, A. B., *Trends in Residential Segregation by Race: 1960-1980*, Report 3627, Washington: Urban Institute, 1986.
- Newburger, H., *Recent Evidence on Discrimination in Housing*, Washington: U.S. Department of Housing and Urban Development, 1984.
- Schnare, A. B., "Trends in Residential Segregation by Race: 1960-1970," *Journal of Urban Economics*, May 1980, 7, 293-301.
- Schwartz, M. A., *Trends in White Attitudes Toward Negroes*, Report No. 119, Chicago: National Opinion Research Center, 1967.
- Sen, A. K., *Collective Choice and Social Welfare*, San Francisco: Holden-Day, 1970.
- Sorensen, A., Taeuber, K. and Hollingsworth, L., "Indexes of Racial Residential Segregation for 109 Cities in the United States, 1940 to 1970," *Sociological Focus*, April 1975, 8, 125-33.
- Taeuber, K., *Appendix to A Decent Home: A Report on the Continued Failure of the Federal Government to Provide Equal Housing Opportunity*, Washington: Citizens Commission on Civil Rights, 1983.
- Viorst, M., *Fire in the Streets: America in the 1960s*, New York: Simon and Schuster, 1979.
- Wienk, R. et al., *Measuring Racial Discrimination in American Housing Markets: The Housing Market Practices Survey*, Washington: U.S. Department of Housing and Urban Development, 1979.
- U.S. Commission on Civil Rights, *A Sheltered Crisis: The State of Fair Housing in the Eighties*, Washington: USGPO, 1983.

# UNCERTAINTY IN MACROECONOMICS<sup>†</sup>

## Uncertainty Across Models

By CHRISTOPHER A. SIMS\*

Everyone recognizes—indeed, it is a stale joke—that economists and economic models are likely to offer a variety of different conclusions on any given policy issue. In the face of conflicting formal models, our profession responds along several lines. One line compares the models according to the “reasonableness” of their implications, according to the judgment of some set of economists, without any attempt at formal inference. I will not discuss this approach further in this paper, but it must be noted with dismay that, at least in macroeconomics and probably more broadly, this approach is more common than the ones I will be discussing.

The lines of approach I discuss are those of (i) weighting together results from a set of models according to measures of their historical forecast accuracy; (ii) selecting the best among a set of models according to a systematic program of specification tests; and (iii) characterizing the range of results generated by the set of models without merging them or selecting from them.

### I. A Point of View

A model for econometric inference that I find enlightening treats it as Bayesian inference in an infinite dimensional topological vector space (TVS). I have discussed the implications of this point of view in general and specifically for distributed lag estimation when lag length is unknown (1971). Recently, A. Ronald Gallant and J. F. Monahan (1985) have applied it to the non-

linear regression problem. There is not space here to explain this approach in detail, but readers interested in the broader foundations for the specific discussion that follows might consult the references.

This model for inference shows that econometricians’ propensity to consider several finitely parameterized models, often without being able to choose clearly among them, can be explained as part of a reasonable strategy for inference in a large parameter space. In fact, estimation strategies that consider a range of finitely parameterized models, choosing among them or averaging them according to how well they fit and according to how complex they are (penalizing more complex models), are fully as general as explicitly nonparametric approaches. (Examples of explicitly nonparametric approaches are most frequency domain estimation in time series and the kernel regression estimators surveyed by Herman Bierens, 1987.)

While this viewpoint justifies much of what econometricians ordinarily do, it does not justify the common practice of emerging from contact with the data with a single model, as if probability statements conditioned on the truth of that model captured all the real uncertainty. It is possible to deal with uncertainty across models, and I discuss below three approaches to doing so.

### II. Reporting for a Range of Models Without Evaluating Them

Edward Leamer (1987), indicates that there are some types of uncertainty we should rightly recoil from quantifying as probabilities. For this type of uncertainty, he suggests, we should be willing to simply postulate a range of values for the uncertain parameter, then keep track of the range of

<sup>†</sup>*Discussants:* Edward E. Leamer, UCLA; Michael Woodford, University of Chicago.

\*University of Minnesota, Minneapolis, MN 55409, and Federal Reserve Bank of Minneapolis.

implications that can emerge from this range of parameter settings when the model confronts the data.

Let me paraphrase Leamer. (Since the paraphrase brings out my doubts about his suggestion, it may well distort it.) If what we don't know about the world is summarized in a list of unknown parameters and random variables, he suggests that for the random variables and some of the parameters (because he is a variety of Bayesian) we postulate prior probability distributions. For other parameters, about which our beliefs are too nonquantitative and imprecise, we make no attempt at postulating a prior. Our analysis is then, in principle, repeated conditioning on every possible setting of these nonrandom parameters, but we attempt to avoid doing that much work by discovering the extremes of the range of results generated by different settings of these nonrandom parameters.

Notice that an ordinary Bayesian analysis can be described as beginning with a determination of model results conditioned on each setting of the full list of parameters, then proceeding to a weighting together of results using the prior *and the sample likelihood* to form weights. If there is no specific decision or function of the parameters as a focus of interest, the Bayesian analysis summarizes the shape of the product of the likelihood function with the prior probability density function (i.e., of the posterior p.d.f.). What Leamer suggests differs from this (which he describes as a "proper" Bayesian procedure, 1987, p. 13-14) only in that he seems to insist on suppressing the weighting of the results in the likelihood. A full likelihood, over the nonrandom as well as the random parameters, of course contains implicitly all the likelihoods conditional on the nonrandom parameters, which is what Leamer indicates be reported. In practical inference problems, it is often a great simplification to exclude regions of the parameter space that have such low likelihood that no plausible set of prior beliefs could make them a posteriori likely. Leamer's exposition of his idea suggests that what he proposes is easier than "proper" Bayesian analysis. But a proper Bayesian who is willing to use flat or otherwise conventional priors for report-

ing purposes (like me) does not have tortuous work to do in eliciting a prior, and he does have the convenience of being able to omit sensitivity analyses in regions of the parameter space that the data clearly rule out.

Nonetheless, Leamer's methods ought to be more widely used. There are three reasons. The first two I pointed out elsewhere (1987)—sometimes results conditional on parameter settings are cheap to compute, while assessing the prior and the likelihood are not cheap; and sometimes we are reporting results that are sensitive to differences in prior beliefs that we might expect to obtain across users of our analysis. Another reason is that the usual econometric style is to make assumptions that eliminate dimensions of uncertainty about which the data contain little information. Though textbook treatments provide no justification for it, econometricians often speak of "identifying a model" by "making identifying assumptions." By this they often mean making arbitrary assumptions about parameters whose values the data can in any case not determine. Often these dimensions of uncertainty are quite important to policy conclusions, yet the effect of real uncertainty of this type is not made explicit in announcing the conclusions of the empirical study. Leamer's methods might be a path of least resistance away from econometricians' conventional arrogance in putting forth the implications of their econometric models. A proper Bayesian approach, using flat or conventional priors, could produce results as easily, but in these situations, where the data provide little information about the parameters in question, there is not much difference between Leamer's proposal and a properly interpreted flat-prior Bayesian data analysis in which the likelihood is described or summarized.

### III. Specification Testing Schemes

Specification testing schemes are in a sense the opposite extreme from the sensitivity analysis approach discussed above. While the sensitivity analysis approach pays no explicit attention to measures of fit, simply displaying the range of results with various models, the specification testing approach

focuses all attention on models' fit to the data, eventually emerging with only one or a small number of models that fit acceptably. As already noted, focusing attention on a small number of models that fit well can be justified as part of an approximate Bayesian approach to dealing with an infinite dimensional parameter space.

David Hendry's work (summarized, 1987) has improved practical implementations of sets of tests of fit, with the "encompassing" idea pulling specification testing somewhat closer to a Bayesian ideal. Classical fit tests of different models often involve different statistics computed from the same data. By insisting that a good model should be able to explain the results of test of other less good models as well as to pass tests of its own fit, the encompassing principle leads to procedures that are close to comparing likelihoods across models.

But specification testing schemes that are not grounded in a Bayesian framework confront several pitfalls. In situations with many competing models or highly multivariate or nonlinear models, classical inference may provide little guidance on how to interpret an array of tests of fit. If we are forming opinions about the location of a parameter in an infinite dimensional space, we will need to specify a sequence of expanding subspaces or compact subsets of the space on which probability concentrates, and prior probability weights on these component subspaces or models. If this is not done, the results can be unreasonable or bizarre. In the time-series literature, classical tests for length of lag have been recognized as a mistaken approach to choosing among models differing only by length of lag, for example. Such tests, if conducted at a conventional significance level that does not change with sample size, will not lead to convergence on the true lag length as sample size increases, even if the data are generated by a finite lag length model. A Bayesian analysis shows that there should be a systematic change in the significance level of tests with sample size, and this conclusion does not depend much on the form of the prior in large samples. Leamer's (1978) modification of the regression  $F$  test applies this same idea. Without appeal to this kind of Bayesian reasoning, it will inevi-

tably be hard to explain how to choose between simpler and more complicated models. Hendry's summary refers to "parsimony" (the good property of a model with a small number of parameters) in passing, but gives it no systematic discussion. In highly multivariate systems finding a systematic way to give appropriate credit to a model for parsimony is essential, and without a Bayesian underpinning, rules for giving such weight appear arbitrary and mysterious.

Another problem with specification testing is that when applied carefully it will ordinarily lead to a conclusion that several competing models are more or less consistent with the data. Since this is an unsatisfying conclusion, there is a tendency to set up the procedure, or at least to describe it *ex post*, so that it yields a unique best model. Indeed Hendry (p. 42) appears to claim never to have encountered a situation in which there was no unique best model. He also asserts (sec. 4) that the process by which a model is arrived at (i.e., the class of models initially considered, how it was sorted through in interaction with the data) need not be reported. He is confident that after a model is "discovered," a subsequent period of "justification" will be adequate protection against overfitting and spurious results. This position is just not tenable in econometrics. Very seldom is a model on a scale big enough to be of practical use formulated years in advance of its actual application. Nearly always the sample period, used to formulate the model, contains vastly more data than any subsequent period that can be used for justification. In cross-section applications, there is often no prospect of a subsequent period of justification before the model has to be applied. Furthermore, I think there is no example (somebody will undoubtedly find one for me, given that I've put it this strongly) of an econometric model, useful in practice, that a well-trained economist could not show to contain several simplifying assumptions that are clearly false, conceivably quantitatively important, and critical to actual use of the model. Whereas Leamer's ideas on sensitivity analysis point the researcher toward locating dimensions of uncertainty on which the data are uninformative, the specification testing framework points in the opposite di-

rection, reinforcing our discipline's tendency to hide its uncertainty.

#### IV. Combining Models According to Forecast Performance

If we have a vector  $z$  of  $k$  variables to use in predicting the variable  $y$ , econometricians tend to generate a linear regression model projecting  $y(t)$  on  $z(t)$ . If  $z(t)$  happens to be a vector of forecasts of  $y(t)$  generated by  $k$  different models, the main specialization is that, assuming all the models take adequate account of serial correlation, there is no need to consider models using lagged as well as current values of  $z$  on the right-hand side. In macroeconomic policy analysis, it is common for there to be available projections of the effects of policy actions based on several large, complicated models. It has been a recurrent idea that one ought to be able to use regression methods to construct optimal weighting schemes for the results from the models.

I see three categories of difficulty with this idea. First, the method amounts to constructing a single forecasting model out of the original  $k$  models, while ignoring all the internal structure of the  $k$  models. This is essentially the same point as the encompassing idea in Hendry's specification testing scheme. If we have two models, one a regression of  $y$  on  $x$  and  $z$ , the other a regression of  $y$  on  $x$  alone, we might see in a moderate-sized sample that the first model has a worse overall forecasting record because early in the sample the estimate of the coefficient of  $z$  was ill-determined, but that now the coefficient is sharply determined and large. We would then expect that ignoring  $z$  in forecasting is a mistake and the first model is the better of the two. No stationary regression of  $y$  on historical forecasts from the two models could show us this. A regression with time-varying coefficients might come close to doing so, but such a model would be more complicated than the original pair of models. If it is possible, it must be better to look at why forecasts differ among the models in coming to conclusions about how to judge their current forecasts.

This point becomes much more important when we confront the surprisingly common

naive idea that forecast accuracy is the ultimate test of a model's "truth," and that the best forecasting model therefore will yield the best policy conclusions. When we are considering several models, each of which represents considerable thoughtful labor and has been used for forecasting, it is unlikely that the main differences among them are well described by ranking them according to their distance from the truth. More likely they each represent a judicious compromise between parameter parsimony and honest representation of uncertainty. If  $z$  has varied little in the past, good forecasting models are very likely to set the effects of  $z$  to zero or to some a priori reasonable guessed value. If our current policy problem is to predict the effects of a large change in  $z$ , we might well do better with models that do not forecast well because they have used sample information in determining the coefficient of  $z$ .

Finally, econometricians are only beginning to discover how complicated the joint distribution of macroeconomic time-series forecast errors really is. R. F. Engle's (1982) work on ARCH models has shown that time variation in the scale of forecast errors is important. My own recent work (1988) with a quarterly forecasting model implies that nonnormality and therefore nonlinearity of regression relations among forecasting errors is also important. Indeed, nonnormality and nonstationarity interact, so that when both are allowed for simultaneously, each appears more important. Covariances among forecast errors in different variables shift size and even sign between subperiods of the postwar period in the United States, and the shifts are statistically significant under a stationary null hypothesis. Furthermore, there are recurrent episodes of forecast errors much larger than would be consistent with a stationary Gaussian model. If these facts are not allowed for in analyzing models' historical forecasting records, results are likely to be anomalous or disappointing.

#### V. Conclusion

This paper has been critical of some of the claims made for each of the methodological lines it discusses. The generality of nonparametric econometrics relative to finitely pa-



parameterized models is illusory. Systematic sensitivity analysis is a prescription for unnecessary complexity if it completely forswears using measures of fit to eliminate models. Specification testing tends to reinforce the myth that econometric analysis should ordinarily emerge with a single model in which the data can determine the answers to whatever questions about behavior we may ask. And use of forecasting records to rank or weight together models may for several reasons give disappointing or misleading results.

But I should close by restating that each of these methodological lines is used less commonly than it should be. Each represents a step away from the widespread pattern of econometric research in which results are reported as if uncertainty across models did not exist.

#### REFERENCES

- Bewley, Truman, *Advances in Econometrics Fifth World Congress*, Cambridge: Cambridge University Press, 1987.
- Bierens, Herman J., "Kernel Estimates of Regression Functions," in T. Bewley, ed., *Advances in Econometrics Fifth World Congress*, Cambridge, 1987, 99-144.
- Engle, R. F., "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of U.K. Inflation," *Econometrica*, July 1982, 50, 987-1008.
- Gallant, A. Ronald and Monahan, J. F., "Bayesian Estimation of the Fourier Flexible Form," *Journal of Econometrics*, October/November 1985, 30, 171-220.
- Hendry, David, "Econometric Methodology: A Personal Perspective," in T. Bewley, ed., *Advances in Econometrics Fifth World Congress*, Cambridge, 1987, 29-48.
- Leamer, Edward E., *Specification Searches*, New York: Wiley & Sons, 1978.
- \_\_\_\_\_, "Econometric Metaphors," in T. Bewley, ed., *Advances in Econometrics Fifth World Congress*, Cambridge, 1987, 1-28.
- Sims, Christopher A., "Distributed Lag Estimation when the Parameter Space is Explicitly Infinite Dimensional," *Annals of Mathematical Statistics*, Vol. 42, 1971, 1622-36.
- \_\_\_\_\_, "Making Economics Credible," in T. Bewley, ed., *Advances in Econometrics Fifth World Congress*, Cambridge, 1987, 49-61.
- \_\_\_\_\_, "A Nine Variable Probabilistic Macroeconomic Forecasting Model," unpublished, 1988.

# The Fate of Systems With "Adaptive" Expectations

By ALBERT MARCET AND THOMAS J. SARGENT\*

This paper provides an informal summary of recent results from the literature on convergence of least squares learning schemes to rational expectations equilibria in dynamic self-referential environments. This literature answers the following questions. 1) How will a system eventually behave in which agents make forecasts with recursively updated least squares estimates of vector autoregressions and in which agents' forecasts influence the law of motion of the whole system?<sup>1</sup> 2) In systems with multiple rational expectations equilibria, how does convergence of least squares learning provide a stability criterion for selecting among equilibria? 3) Do least squares learning schemes suggest new methods for computing rational expectations equilibria? 4) Do self-referential systems with least squares learning offer a promising set of econometric methods that are an alternative to rational expectations econometrics? 5) How do systems with least squares learning relate to informal descriptions which model learning in rational expectations models by iterating on the operator mapping perceived laws of motion into actual laws of motion?<sup>2</sup>

We summarize this literature from the perspective of our own work, which has found a close connection between the operator mentioned in question 5, and the circumstances

governing convergence to rational expectations equilibria of the systems under study in question 1.<sup>3</sup>

Consider a class of self-referential dynamic models designed to restrict a stochastic process for a vector  $z_t$  of state variables. The model states that  $z_t$  obeys the law

$$(1) \quad z_t = f(z_{t-1}, \varepsilon_t),$$

where  $\varepsilon_t$  is a vector white noise. Economic reasoning makes the function  $f$  itself a functional of two matrices,  $\beta_a$  and  $\beta_b$ , that summarize the beliefs of two differentially informed classes of agents, each of which lives within the model, and each of whose decisions impinge on the law of motion  $f$ . In particular, we suppose that  $\beta_a$  (or  $\beta_b$ ) characterizes a linear rule by which agents of class  $a$  (or  $b$ ) forecast the uncontrollable variables in their objective functions and constraints, forecasts of which are needed to make decisions. The agents of class  $a$  and  $b$  are each typically assumed to observe only a subset of the state vector  $z_t$ , say  $z_{at}$  or  $z_{bt}$ , respectively, and must base their forecasts on this observed subset. Agents' decisions, which are based partly on their perceptions,  $\beta_a$  and  $\beta_b$ , influence the actual law of motion of the state vector  $z_t$  given in (1). We express this dependence via the functional

$$(2) \quad f = T(\beta_a, \beta_b).$$

We have in mind models in which the  $T$  operator is derived using decision theory that attributes to agents no uncertainty around the perceptions  $(\beta_a, \beta_b)$ .

Given an arbitrary pair  $(\beta_a, \beta_b)$ , it is possible to calculate from (1) and (2) the perceptions of agents  $a$  and  $b$ , respectively.

<sup>3</sup>See our papers (1987a, b, c). We rely heavily on technical results of Ljung which are extensively applied by Ljung and Söderström.

\*Carnegie-Mellon University, Pittsburgh, PA 15213, and Hoover Institution, Stanford University, Stanford, CA 94305 and Federal Reserve Bank of Minneapolis, respectively. Sargent's research was supported by NSF grant SES-85-08935 to the University of Minnesota.

<sup>1</sup>In the language of the literature on control, this is an example of an "adaptive" system. See Lennart Ljung and Torsten Söderström (1983).

<sup>2</sup>Stephen DeCanio (1979) and George Evans (1983, 1985) used this operator to study the stability of rational expectations equilibria. Finn Kydland and Edward Prescott (1977) used iterations on such an operator to describe the forces propelling their inflation-unemployment game to a (suboptimal) time consistent equilibrium.

which would be optimal (i.e., yield least squares forecasts) given their information sets. We denote these perceptions  $S_a(\beta)$ ,  $S_b(\beta)$ , respectively, where  $\beta = (\beta_a, \beta_b)$ . The model thus induces a mapping from a pair  $(\beta_a, \beta_b)$  of arbitrary perceptions to a pair  $S_a(\beta)$ ,  $S_b(\beta)$  of optimal perceptions (regressions). We use the standard *Definition: A rational expectations equilibrium is a set of perceptions  $(\beta_a, \beta_b)$  that satisfy  $(\beta_a, \beta_b) = (S_a(\beta), S_b(\beta))$* . Note that this definition takes as given whatever information discrepancies the analyst has built into the model. We denote a rational expectations equilibrium as  $\beta_f = (\beta_{af}, \beta_{bf})$ .

A variety of models exhibit the self-referential structure depicted in (1) and (2). Among these are Robert Lucas and Edward Prescott's (1971) model of investment under uncertainty, Margaret Bray's (1982) model of equilibrium with informed and uninformed traders, Robert Townsend's (1983) model of forecasting the forecasts of others, and a version of Finn Kydland and Prescott's model of a Nash feedback equilibrium between a government and a competitive private sector in an inflation-unemployment game.

Superimposed on top of (1) and (2), there is a class of least squares learning models that are constructed as follows. Suppose that at time  $t$ , agents of class  $j (= a, b)$  form their perceptions by constructing an estimator by some version of least squares based on whatever components of  $z_t$  they have observed through time  $t-1$ . Letting  $R_{jt}$  be the sample moment matrix of the information available to agents of type  $j$  at time  $t$ , we can represent least squares recursively as

$$(3a) \quad \beta_{jt} = s_j(\beta_{jt-1}, z_{t-1}, R_{jt-1}), \quad j = a, b;$$

$$(3b) \quad R_{jt} = r_j(R_{jt-1}, z_{t-1}), \quad j = a, b.$$

The system  $(z_t, \beta_t)$  is assumed to evolve according to (3) and

$$(4) \quad z_t = f_t(z_{t-1}, \varepsilon_t)$$

where

$$(5) \quad f_t = T(\beta_{at}, \beta_{bt})$$

In (5),  $T$  is the same operator that appears in (2). Because this operator is typically derived under the assumption that  $\beta_j$  is known for sure, the model consisting of (3), (4), (5) builds in irrationality.

The literature has focused on the behavior of the estimators  $\beta_{jt}$  associated with (3), (4), (5) as time passes without limit. Research has aimed to characterize conditions under which  $(\beta_{at}, \beta_{bt})$  converge to a rational expectations equilibrium,  $(\beta_{af}, \beta_{bf})$ . Main findings of the literature are

I. If  $\{\beta_{at}, \beta_{bt}\}_{t=0}^{\infty}$  converges, it converges to a rational expectations equilibrium. More precisely, if  $(\hat{\beta}_a, \hat{\beta}_b) \neq (\beta_{af}, \beta_{bf})$ , then  $\text{Prob}\{(\beta_{at}, \beta_{bt}) \rightarrow (\hat{\beta}_a, \hat{\beta}_b)\} = 0$ . Versions of this proposition are proved by Bray, Bray and Eugene Savin (1986), and our paper (1987a).

II. The local stability of the learning system about a rational expectations equilibrium  $(\beta_{af}, \beta_{bf})$  is determined by an associated differential equation system

$$(6) \quad \frac{d}{dt} \begin{pmatrix} \beta_a \\ \beta_b \end{pmatrix} = \begin{bmatrix} S_a(\beta) - \beta_a \\ S_b(\beta) - \beta_b \end{bmatrix} = S(\beta) - \beta.$$

To study local convergence, determine the eigenvalues of the Jacobian matrix associated with the right side of (6) at a fixed point  $(\beta_{af}, \beta_{bf})$ . If there is one or more eigenvalue with strictly positive real part, then  $\text{Prob}\{(\beta_{at}, \beta_{bt}) \rightarrow (\beta_{af}, \beta_{bf})\} = 0$ . If all of these eigenvalues are less than zero in real part, then there exists a set containing  $(\beta_{af}, \beta_{bf})$  such that if  $(\beta_{at}, \beta_{bt})$  remains within this set, then  $\text{Prob}\{(\beta_{at}, \beta_{bt}) \rightarrow (\beta_{af}, \beta_{bf})\} = 1$ .

III. Global convergence is governed by the behavior of a larger differential equation system, namely,

$$(7) \quad \frac{d}{dt} \begin{bmatrix} \beta'_a \\ \beta'_b \\ R_a \\ R_b \end{bmatrix} = \begin{bmatrix} R_a^{-1} M_a(\beta) [S_a(\beta) - \beta_a]' \\ R_b^{-1} M_b(\beta) [S_b(\beta) - \beta_b]' \\ M_a(\beta) - R_a \\ M_b(\beta) - R_b \end{bmatrix}$$

In (7),  $M_{z_j}(\beta)$  is the matrix  $Ez_{jt}z'_{jt}$  calculated from the stationary distribution of (1), (2) at a fixed value of  $\beta = (\beta_a, \beta_b)$ . Note that a stationary point of (7) is a rational expectations equilibrium, with  $\beta = S(\beta)$ ,  $R_a = M_{z_a}(\beta)$ ,  $R_b = M_{z_b}(\beta)$ . Sufficient conditions for global convergence have been obtained by restricting the least squares algorithm (3a)–(3b) to require that  $(\beta_t, R_t) \equiv (\beta_{at}, \beta_{bt}, R_{at}, R_{bt})$  lie within a bounded set  $D$  that contains a stationary point of (7). This is achieved by naming a set  $D$ , and modifying the least squares algorithm (3) to ignore observations that threaten to drive  $(\beta_t, R_t)$  outside of  $D$ . The sufficient conditions for global convergence require that the set  $D$  lie within the domain of attraction of the stationary point of (7), and that the trajectories of (7) point toward the interior of  $D$  at and near the boundary of  $D$ .

Result I informs us that among time invariant representations that describe the eventual behavior of the learning system (3), (4), (5), there are only rational expectations equilibria. Thus, least squares learning models of the form (3), (4), (5) seem not to provide any basis for fitting time invariant models that are not rational expectations models.

Results II and III clarify the relationship between the literatures mentioned in questions 1 and 5 above. On the one hand, DeCanio, Evans (1983, 1985), and Kydland-Prescott described processes which amounted to iterating on the  $S$  operator.

$$(8) \quad \beta_k = S(\beta_{k-1}).$$

On the other hand, numerical solutions of the differential equation (6), which the least squares learning algorithm actually approximates, can be obtained via Euler's method:

$$(9) \quad \beta_k = \beta_{k-1} + \gamma[S(\beta_{k-1}) - \beta_{k-1}]$$

for  $\gamma > 0$ ,  $\gamma$  small. Note that (8) is (9) with  $\gamma = 1$ . Convergence of (8) to a fixed point implies convergence of (9) with  $0 < \gamma < 1$ , but not vice versa. If the eigenvalues of the Jacobian of  $S$  evaluated at  $\beta_f$  are less than  $-1$  in real part, then (8) will not converge

even though there exists a  $\gamma > 0$  for which (9) converges.

These observations imply that iterations (8) give too pessimistic a criterion for convergence of least squares learning schemes. They also suggest that algorithms for computing a rational expectations equilibrium superior to (8) can be devised. These could be based on numerical solution of either (6) or (7) (for example, as given by (9)), or else by computing a long stochastic simulation of the learning system (3), (4), (5) itself. It is remarkable that this naive learning system has local stability characteristics that are superior to those associated with the algorithm (8), which pioneering work on computing rational expectations equilibria sometimes focused on.

We have computed the  $S$  mapping for a variety of examples, and have used it to interpret the stability conditions obtained in previous work including that by Bray, Bray-Savin, and C. Fourgeaud, C. Gourieroux, and J. Pradel (1986). For many models, the differential equations (6) and (7) assure both local and global convergence for a range of "reasonable" values for the economic parameters that determine  $S$ .<sup>4</sup>

Michael Woodford (1986), Evans (1983, 1985) and ourselves (1987b, c) have discussed using the local stability of the  $S$  mapping described in II as a "correspondence principle" device for selecting equilibria. Woodford shows how stationary sunspot equilibria in an overlapping generations monetary model are stable under least squares learning. Evans (1987) and ourselves (1987b, c) describe examples in which least squares learning schemes fail to converge to "bubble" rational expectations equilibria. The "bubble" equilibria tossed out by our stability analysis are inefficient (they are bad "Laffer curve" equilibria), while the sunspot equilibria retained in Woodford's setting are good in the sense of being conditionally Pareto optimal.

<sup>4</sup>We (1987a) show that the differential equation system (7) assumes the same form for the models both of Bray-Savin and of Fourgeaud et al.

These uses of least squares learning as a selection criterion are not entirely robust with respect to alternative specifications of the learning scheme. James Jordan (1986) has produced a setting in which he is able to produce an adjusted least squares learning scheme capable of converging to any of the rational expectations equilibria of his model. Moreover, as Bray and David Kreps (1987) have emphasized, least squares learning schemes are irrational (for example, they embody a Bayesian prior that is inconsistent with the law of motion (4)–(5)). The above uses of least squares learning as a stability criterion invoke a selection criterion that is drawn from outside the model environment.

It is open and problematic whether the learning system (3), (4), (5) can ever be expected to yield econometric models that can be applied. The extent to which system (3)–(5) deviates from a rational expectations equilibrium is determined by initial conditions for  $(\beta_a, \beta_b, R_a, R_b)$ , which must be estimated as parameters. Results II and III probably imply that these parameters cannot be consistently estimated, because the tail of the stochastic process for  $z_t$  does not depend on them. Standard minimum norm estimators could still be applied to this system. We are unaware of such applications, or of characterizations of the statistical properties of such estimators.

This literature provides mixed comfort. It is remarkable that the “adaptive” least squares learning schemes are attracted to rational expectations equilibria, and that, naive and backward-looking as they are, they provide promising leads on superior ways for us economists to compute rational expectations equilibria. It is also comforting that these adaptive mechanisms seem not to be attracted to “bad” bubble equilibria as limit points. Heuristically, the reason that these bad equilibria are excluded as limit points seems to be that to support them requires extraordinary foresight on the parts of agents, and that the adaptive schemes analyzed cannot accommodate such foresight. This inability to speculate leads to good outcomes in some situations, as in the Woodford and our cited examples, but leads to bad outcomes in versions of the Kydland-

Prescott inflation-unemployment example. In that example, our analysis would lead us to expect that an adaptive game between a government and private sector, each of which estimates and each period updates an econometric model  $(\beta_a, \text{ and } \beta_b)$  via least squares, would approach the Nash-feedback equilibrium, which is suboptimal. That suboptimality is rooted in the adaptive behavior of both government and private agents.

## REFERENCES

- Bray, Margaret, “Learning, Estimation, and the Stability of Rational Expectations,” *Journal of Economic Theory*, April 1982, 26, 318–39.
- and Kreps, David M., “Rational Learning and Rational Expectations,” in George Feiwel, ed., *Arrow and the Ascent of Modern Economic Theory*, New York: New York University Press, 1987, 597–625.
- and Savin, N. E., “Rational Expectations Equilibria, Learning and Model Specification,” *Econometrica*, September 1986, 54, 1129–60.
- DeCanio, Stephen J., “Rational Expectations and Learning from Experience,” *Quarterly Journal of Economics*, February 1979, 93, 47–57.
- Evans, George, “Expectational Stability and the Multiple Equilibria Problem in Linear Rational Expectations Models,” *Quarterly Journal of Economics*, November 1985, 100, 1217–34.
- , “The Stability of Rational Expectations in Macroeconomic Models,” in Roman Frydman and Edmund S. Phelps, eds., *Individual Forecasting and Aggregate Outcomes: “Rational Expectations” Examined*, New York: Cambridge University Press, 1983.
- , “The Fragility of Sunspots and Bubbles,” manuscript, Stanford University, 1987.
- Fourgeaud, C., Gouriou, C. and Pradel, J., “Learning Procedure and Convergence to Rationality,” *Econometrica*, July 1986, 54, 845–68.
- Jordan, James S., “Convergence to Rational Expectations in a Stationary Linear Game,” manuscript, University of Min-

- nesota, May 1986.
- Kydland, Finn E. and Prescott, Edward C., "Rules Rather Than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, 85, 473-92.
- Ljung, Lennart, "Analysis of Recursive Stochastic Algorithms," *I.E.E.E. Transactions of Automatic Control*, Vol. AC-22, 1977, 551-57.
- \_\_\_\_\_ and Söderström, Torsten, *Theory and Practice of Recursive Identification*, Cambridge: MIT Press, 1983.
- Lucas, R. E., Jr. and Prescott, E. C., "Investment Under Uncertainty," *Econometrica*, September 1971, 39, 659-81.
- Marcet, Albert and Sargent, Thomas J., (1987a) "Convergence of Least Squares Learning in Environments with Hidden State Variables and Private Information," manuscript, October 1987.
- \_\_\_\_\_ and \_\_\_\_\_ (1987b) "Convergence of Least Squares Learning Mechanisms in Self-Referential Linear Stochastic Models," manuscript, October 1987.
- \_\_\_\_\_ and \_\_\_\_\_ (1987c) "Least Squares Learning and the Dynamics of Hyperinflation," manuscript, June 1987.
- Townsend, R. M., "Forecasting the Forecasts of Others," *Journal of Political Economy*, August 1983, 91, 546-88.
- Woodford, Michael, "Learning to Believe in Sunspots," C.V. Starr Center Working Paper No. 86-16, New York University, June 1986.

# Consumption: Beyond Certainty Equivalence

By OLIVIER JEAN BLANCHARD AND N. GREGORY MANKIW\*

Twenty years ago, it was standard practice in describing macroeconomic behavior to build theoretical models assuming all current and future variables were known with certainty. When models were applied to the data, the only concession to the presence of uncertainty was more often than not the introduction of an unexplained error term in the regression.

Ten years ago, under injunctions to "take uncertainty seriously," macroeconomists started introducing uncertainty explicitly at the model-building stage. Much theoretical and empirical progress has been made, most of it under assumptions yielding certainty equivalence, or the property that optimal behavior depended only on expectations of other variables, and not on their higher moments.

Certainty equivalence yields convenient characterizations of behavior. Assumptions yielding certainty equivalence, namely that constraints are linear and objective functions quadratic, are however quite stringent and, in most contexts, highly implausible. Recent research has attempted to go beyond certainty equivalence, and to characterize behavior under more appealing assumptions. It is a difficult endeavor, both theoretically and empirically. In most cases, closed-form solutions are unavailable and one only gets glimpses into the nature of the solution. In most cases also, the decision rule depends on higher moments of the exogenous variables, about which little hard evidence is available, making empirical implementation perilous. Nevertheless, much has been learned; in this paper, we present recent developments on the consumption front.<sup>1</sup>

\*MIT, Cambridge, MA 02139 and NBER, and Harvard University, Cambridge, MA 02138 and NBER, respectively.

<sup>1</sup>Developments on other fronts would also warrant a report. See, in particular, Giuseppe Bertola (1987) for an analysis of the interaction of uncertainty and irreversibility in determining investment.

## I. Consumption under Certainty Equivalence

Consider the decision problem of a consumer who maximizes at time  $t$ :

$$(1) \quad E \left[ \sum_{i=0}^{T-t} U(c_{t+i}) | I_t \right]$$

subject to  $A_{t+i+1} = A_{t+i} + Y_{t+i} - C_{t+i}$ ;

$$A_t \text{ given; } A_{T+1} = 0.$$

For simplicity, we assume that both the interest rate and the discount rate are equal to zero.  $A_t$  is wealth,  $Y_t$  labor income. The only source of uncertainty is labor income, which is random.

If utility is quadratic, the set of first-order conditions is

$$(2) \quad E(C_{t+i} | I_t) = C_t, \quad \text{for } i = 1, \dots, T-t$$

and the solution to the maximization problem has the familiar form

$$(3) \quad C_t = (1/(T-t+1)) \times \left( A_t + \sum_{i=0}^{T-t} E[Y_{t+i} | I_t] \right).$$

Consumption is a linear function of initial wealth and the present value of expected future income. Higher moments of income do not matter. The marginal propensity to consume out of total wealth is equal to the inverse of remaining number of years.

The assumption of quadratic utility is crucial to derive the "certainty equivalence" consumption function (3) in the presence of uncertain labor income. Yet, quadratic utility is an unappealing way of describing consumers' behavior towards risk. It implies increasing absolute risk aversion, a willingness to pay more to avoid a given bet as wealth increases. Introspection and casual evidence

suggest that this is a poor description of behavior under uncertainty.

Simple utility functions with more plausible properties towards risk are available, of course. Two such functions are the exponential and the isoelastic utility functions. Yet, in the presence of risky labor income, neither yields certainty equivalence. Indeed, they imply systematic effects of uncertainty on consumption, to which we now turn.

## II. The Slope and Variance of Consumption

If we return to the set of first-order conditions of the maximization problem above, this time without restrictions on utility beyond risk aversion,  $U'' < 0$ , we get

$$(4) \quad E[U'(C_{t+i})|I_t] = U'(C_t),$$

for  $i = 1$  to  $T - 1$ .

Uncertainty affects the first-order conditions, and thus optimal consumption, only if it affects expected marginal utility. If the third derivative of the utility function  $U'''$  is positive, as is true of most plausible utility functions, an increase in uncertainty raises expected marginal utility. Thus to maintain equality in (4), expected future consumption must increase compared to current consumption. Uncertainty leads consumers to defer consumption, to be more prudent. The role of the condition  $U''' > 0$  in generating more prudent behavior in the face of uncertainty was first derived by Hayne Leland (1968) and further analyzed by Agnar Sandmo (1970) and Jacques Drèze and Franco Modigliani (1972).

How strong is the effect of uncertainty on the slope of the consumption path likely to be? Miles Kimball (1987) has shown that, in the same way as the Arrow-Pratt coefficients of risk aversion help study the effects of uncertainty on expected utility, coefficients of absolute and relative prudence help study the effects of uncertainty on expected marginal utility and thus on consumption. In parallel to the Arrow-Pratt coefficients, the coefficient of absolute prudence is defined as  $-U'''/U''$ , and the coefficient of relative prudence as  $-U'''C/U''$ . Constant abso-

lute prudence implies that the increase in consumption required to keep the same level of expected marginal utility in the face of small increase in risk is independent of the initial level of consumption, and a parallel interpretation applies to the coefficient of relative prudence.

In general, there need not be any tight relation between the coefficients of risk aversion and the coefficients of prudence. Conveniently—and perhaps misleadingly—however, the exponential utility function,  $U(C) = -(1/\gamma)\exp(-\gamma C)$ , exhibits both constant absolute risk aversion,  $\gamma$ , and constant absolute prudence, also equal to  $\gamma$ . Similarly, the isoelastic utility function  $U(C) = (1 - \xi)C^{1-\xi}$  exhibits both constant relative risk aversion,  $\xi$  and constant relative prudence,  $(\xi + 1)$ . Thus, under those two specifications, specifying the degree of risk aversion also pins down the degree of prudence.

Equipped with those definitions, we can take a second-order approximation of (4) around  $U'(C_t)$ .<sup>2</sup> Rearranging gives

$$(5) \quad E[(C_{t+i} - C_t)|I_t] \\ = (1/2)aE[(C_{t+i} - C_t)^2|I_t]$$

or, dividing both sides by  $C_t$ :

$$(6) \quad E[((C_{t+i} - C_t)/C_t)|I_t] \\ = (1/2)rE[((C_{t+i} - C_t)/C_t)^2|I_t]$$

where  $a$  and  $r$  are the coefficients of absolute and relative prudence. Equation (5) gives a relation between the slope of the consumption path and the variance of the change in consumption (around zero). Equation (6) gives a relation between the expected growth rate and its variance.

While they still only give a relation between two endogenous variables, those two

<sup>2</sup>As is usual, these formulae can be derived exactly under appropriate assumptions if the consumer's problem is set in continuous time. See Douglas Breeden (1986).



equations show the basic effects of uncertainty on consumption. Uncertainty, by increasing the variance of consumption, leads to a more steeply sloped consumption path. The effect is stronger the larger the coefficient of absolute or relative prudence. And, as increases in uncertainty do not affect the budget constraint, any increase the slope of the consumption path implies a decrease in the initial level of consumption.

### III. Uncertainty and the Consumption Function

To go beyond equations (5) and (6) requires solving for consumption as a function of the income process. This is in general difficult.

The case of exponential utility, of constant absolute prudence, has proven analytically tractable (Kimball and Mankiw, 1987; Ricardo Caballero, 1987). Yet, what makes it tractable, however, also makes it somewhat unattractive. To see why, we consider a simple example, which follows Caballero. Suppose that utility is exponential with exponent  $-\gamma$ , and that labor income follows a random walk with normally distributed innovations with standard deviation  $\sigma$ . It is easy to verify that optimal consumption satisfies

$$(7) \quad E[C_{t+1}|I_t] = C_t + \gamma\sigma^2/2.$$

Using the budget constraint, one can show that the level of  $C_t$  is given by

$$(8) \quad C_t = (1/(T-t+1))A_t + Y_t - (\gamma(T-t)/4)\sigma^2.$$

The slope of the expected consumption path, rather than being equal to zero as under certainty equivalence, is positive and constant; it depends both on the degree of absolute prudence,  $\gamma$ , and the variance of income changes. This in turn implies that the consumption function is the same as under certainty equivalence, except for a negative term which depends on the degree of uncertainty, the degree of prudence, and the horizon.

We can use (7) and (8) to get a feel for magnitudes. If we evaluate the expected rate of growth of consumption at a point where consumption and labor income are roughly equal, equation (7) implies

$$(9) \quad (E[C_{t+1}|I_t] - C_t)/C_t = (\gamma C_t)(\sigma/Y_t)^2/2.$$

Using panel data, Robert Hall and Frederic Mishkin (1982) found that the standard deviation of the change in permanent income was about \$1200; as median household income was about \$12,000 during the period (1972), this finding suggests a value of  $\sigma/Y$  of about .1. The term  $\gamma C_t$  is equal to the coefficient of relative risk aversion. If we assume this coefficient to be equal to 4, then equation (9) implies an expected growth rate of consumption of 2 percent. This number is roughly the same as the growth in aggregate consumption per capita. Since the cross-sectional age-consumption profile is upward sloping, the growth in individual consumption must be at least 2 percent.<sup>3</sup>

Cumulated over many years, such a tilt in the consumption path implies substantially lower consumption at the beginning of life, and thus much higher average wealth. Indeed, and this reveals the unattractive aspect of the assumption of constant absolute prudence, equation (8) can easily generate negative initial consumption as a result of uncertainty. Negative consumption is not ruled out by the exponential utility.

When we turn to more attractive utility functions which do rule out negative consumption, such as isoelastic utility, obtaining closed-form solutions becomes generally impossible. But, from some analytical results (Kimball), and from simulations (Stephen Zeldes, 1984; Robert Barsky, Mankiw, and Zeldes, 1986), we know the consumption function has the following property. Under decreasing absolute prudence, the convenient dichotomy between the effects of ex-

<sup>3</sup>Michael Kuehlwein (1987) studies the relation between the growth and variance of consumption in panel data.

pected income and the effects of uncertainty which is exhibited in (8) disappears.

On the one hand, the impact of uncertainty on consumption depends on the level of wealth. At higher levels of wealth, a larger portion of lifetime income is certain—under our assumption of a constant interest rate—and the variance of the percentage change in consumption decreases. This in turn implies a flatter consumption path. Zeldes shows for example that under isoelastic utility, the consumption path is initially very steep and flattens as wealth accumulates. This effect is very much in accordance with empirical evidence. Laurence Kotlikoff and Lawrence Summers (1981, Figure 1), for example, show that the annual rate of change of consumption for the cohort born in 1910 was over 3 percent from age 18 to age 50, but was 1 percent thereafter.

On the other hand, the marginal propensity to consume depends on the amount of uncertainty. An increase in income decreases the need for precautionary savings, leading to a larger response in consumption than would be the case under certainty equivalence. As a result, consumption can show what appears as excess sensitivity to income movements (Zeldes).

#### IV. Changes in Uncertainty and Movements in Consumption

If uncertainty is an important determinant of the level of consumption, changes in uncertainty can potentially be an important source of fluctuations in consumption.

Measuring changes in individual income uncertainty is difficult given the typically short time series on individuals in panel data. A useful starting point is to look at changes in aggregate income uncertainty. To do so, we computed the standard deviation of  $n$ -period ahead forecasts of GNP by DRI, probably a good proxy for the relevant measure of subjective uncertainty. Each month, in addition to its main forecast, DRI issues a set of two or three alternative forecasts for the next three years. Each forecast is given a probability by DRI. When we computed DRI's subjective uncertainty, three results stood out. 1) At each date, the subjective

standard deviation increases roughly as the square root of the forecast horizon, indicating that the uncertainty about the future level of output increases with the horizon. 2) The subjective standard deviation, three years ahead, fluctuates substantially: it varies between 1.14 percent in 1978, and 2.70 percent in 1981. 3) The level of aggregate uncertainty is small relative to the standard deviation of income uncertainty facing individuals, roughly 17 percent over three years (based on Hall and Mishkin).

This last fact suggests that if all consumers share fluctuations equally, movements in aggregate uncertainty are unlikely to have a large impact on aggregate consumption. But if fluctuations fall more heavily on some individuals, the aggregate effect can be much larger. If we assume for example that all consumers follow equation (8), and that only  $\alpha$  percent of the consumers are subject to the aggregate shocks, it is easy to show that the effect on aggregate consumption is proportional to  $1/\alpha$ . The more concentrated the effect of aggregate fluctuations, the stronger the impact of uncertainty on aggregate consumption.

This impact of changing uncertainty underlies the papers by Barsky et al. and by Kimball-Mankiw. Both emphasize the deviations from Ricardian equivalence caused by the interaction between precautionary saving and idiosyncratic income risk. If taxes vary with income, increases in taxes have, in addition to their direct effect on expected after-tax income, an insurance effect which works in the opposite direction. Barsky et al. use simulations to show, assuming isoelastic utility, that debt finance—a decrease in taxes today financed by higher proportional taxes later—can have a significant impact on consumption. They conclude that, for plausible parameter values, the marginal propensity to consume out of a tax cut is approximately half the marginal propensity to consume out of wealth. Kimball and Mankiw derive analytic results for the case of exponential utility. They show that, if individual income is serially correlated, the initial effect of deficit finance on consumption is stronger, the larger the anticipated length of time to the eventual tax increase. The reason is a simple one: the

longer the deferral, the more uncertain individual income and the higher the insurance effect of future taxes.

### V. Conclusion

While macroeconomists have long understood the behavior of consumers under certainty equivalence, the behavior of consumers with plausible utility functions facing uncertain future income has remained largely a mystery. Recent research has begun to reveal some the properties of optimal consumer behavior under uncertainty. Perhaps most important, this research has taught us that, in many ways, the assumption of certainty equivalence can be highly misleading.

### REFERENCES

- Barsky, Robert B., Mankiw, N. Gregory and Zeldes, Stephen P., "Ricardian Consumers and Keynesian Properties," *American Economic Review*, September 1986, 76, 676-91.
- Bertola, Giuseppe, "Irreversible Investment," mimeo., MIT, November 1987.
- Breeden, Douglas T., "Consumption, Production, Inflation, and Interest Rates: A Synthesis," *Journal of Financial Economics*, May 1986, 16, 3-39.
- Caballero, Ricardo, J., "Consumption and Precautionary Savings: Empirical Implications," mimeo., MIT, April 1987.
- Drèze, Jacques H. and Modigliani, Franco, "Consumption Decisions under Uncertainty," *Journal of Economic Theory*, December 1972, 5, 308-35.
- Hall, Robert E. and Mishkin, Frederic S., "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households," *Econometrica*, March 1982, 50, 461-81.
- Kimball, Miles S., "Essays on Intertemporal Household Choice," unpublished doctoral dissertation, Harvard University, 1987.
- \_\_\_\_\_ and Mankiw, N. Gregory, "Precautionary Saving and the Timing of Taxes," mimeo., Harvard University, February 1987.
- Kotlikoff, Laurence J. and Summers, Lawrence H., "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, 89, 706-32.
- Kuehlwein, Michael K., "Consumption in the Presence of Uncertainty," unpublished doctoral dissertation, MIT, 1987.
- Leland, Hayne E., "Saving and Uncertainty: The Precautionary Demand for Saving," *Quarterly Journal of Economics*, August 1968, 82, 465-73.
- Sandmo, Agnar, "The Effect of Uncertainty on Saving Decisions," *Review of Economic Studies*, July 1970, 37, 353-60.
- Zeldes, Stephen P., "Optimal Consumption with Stochastic Income," unpublished doctoral dissertation, MIT, 1984.

# Macroeconomic Implications of the Information Revolution

By GEORGE M. VON FURSTENBERG AND ESFANDIAR MAASOUMI\*

Surely not all time-series data have been improved nor have all data concepts been clarified progressively in economics. Indeed, any proliferation in the underground economy, a withdrawal of resources from data-gathering agencies, and growing obsolescence of past classification and measurement conventions could have diminished the reliability of macroeconomic aggregates for some purposes. On the whole, however, we take it that data have become so much more substantiated, detailed, timely, and accurate as to support an ongoing information revolution. At the same time, the costs of data access and dissemination, storage, and processing have declined to the point of making vastly more information effectively available and public. Before showing how macroeconomic relations are affected by some of this, a brief illustration is offered of how information can be defined by its consequences for the probabilities assigned at first to an exhaustive list of  $n$  mutually exclusive events. This is done with the concept of simple entropy ( $EN$ ) which is frequently used as an inverse measure of information.

If it had been known beforehand which of the  $n$  events was bound to happen—perhaps because Bayesian learning from the accumulation of uniformly relevant data already had identified the one with sole claim to the “truth”—no further information would be gained from its occurrence. There would be zero entropy (uncertainty) in this limiting case, very far from the maximum entropy of  $\ln(n)$  that would be displayed by a rectangular distribution of  $n$  elements, each with probability,  $p_i > 0$  of  $1/n$ . Normally, however, the degree of simple entropy stays be-

tween these extremes for the measure,

$$(1) \quad EN(p) = \sum_{i=1}^n p_i \ln(1/p_i),$$

$$\sum_{i=1}^n p_i = 1.$$

Then the release of economic data brings news that triggers a learning process about salient elements in the open set of contingent events and about their probability weights and determinants.

If only the probability weights change, so that the probabilities  $p_i$  become  $q_i$  in a given set of prospective events, the information gain (–) or loss (+) is measured by the change in entropy as  $EN(q) - EN(p)$ . The fact that this difference can be either positive or negative shows that the release of data need not reduce uncertainty or reinforce an agent's previously held beliefs. Rather, a learning process may be initiated that has the capacity of not only changing probability weights but also precision and content of prior beliefs. In other words, the size of the weights  $p_i$ , the inverse of the variance attributed to them, and the descriptors of any of the events  $i$  as well as their total number  $n$  may all be affected by news releases. Any change in the frequency, coverage, or quality of these releases will change the way learning and revalidation may proceed.

## I. Implications of Data Improvements

Because of the use of benchmarks and other conventions that provide time averaging and shared reference for several time series, measurement errors in economic variables may rarely follow a stationary, serially independent random process. Nor are such errors likely to be contemporaneously uncorrelated across the series. Nevertheless, it is useful to start with this simple specification

\*Departments of Economics, Indiana University, Bloomington, IN 47405, and University of California, Santa Barbara, CA 93106, respectively.

to show how data improvements could affect the structures estimated over time even when the underlying economic behavior, that is, the behavior conditional upon the "true" variables, does not change at all. Assuming only two observable data series in a linear model with errors in variables, the pair of data observed at time  $t$  ( $Y_t, X_t$ ), is distinguished from the associated latent variables,  $y_t$  and  $x_t$ . The lowercase variables are the ones on which economic behavior is based subject to the disturbances  $e_t$ . With measurement errors  $u_t$  and  $v_t$ , the elementary system to consider is

$$(2) \quad Y_t = y_t + u_t; \quad X_t = x_t + v_t$$

$$\text{and} \quad y_t = a + bx_t + e_t.$$

Substituting observable variables and ignoring the intercept yields the familiar single-equation model (see Dennis Aigner et al., 1984, pp. 1324–25),

$$(3) \quad Y_t = bX_t + u_t + e_t - bv_t.$$

Its variance-covariance structure ( $S$ ) is

$$(4) \quad S_{YY} = bS_{YX} + S_{uu} + S_{ee},$$

$$\text{where} \quad S_{YX} = bS_{XX} - bS_{vv},$$

$$\text{and} \quad S_{XX} = S_{xx} + S_{vv}.$$

Without more information than the data  $(Y_t, X_t)$  can provide, system (4) cannot be solved for the structural parameter,

$$(5) \quad b = S_{yx}/S_{xx} = S_{YX}/(S_{XX} - S_{vv}),$$

because it contains 3 equations with 5 unknowns ( $b$  and 4 latent variances). This may lead to bias being accepted in estimating the system by OLS as

$$(6) \quad Y_t = BX_t + W_t, \quad B = S_{YX}/S_{XX}.$$

With one additional bit of information, one may now be able to tell a great deal from changes ( $d$ ) in the variances of the observed

variables. For instance, if there is reason to believe that  $S_{XX}$  has declined over time solely on account of improvements in data quality, it follows that  $dB$  will have the same sign as  $S_{YX}$  and hence  $B$ :

$$(7) \quad dS_{XX} = dS_{vv} < 0; \quad dS_{YX} = 0;$$

$$dS_{YY} = dS_{uu} + dS_{ee};$$

$$db = 0; \quad dB = -S_{YX}dS_{vv}/(S_{XX})^2.$$

The second implication in (7), covariance stationarity, can be used to test whether the first assumption is appropriate so that  $b$  does not change. Its biased estimate,  $B$ , increases asymptotically in absolute value from 0 toward  $b$  as the noise in  $X$  declines toward 0. In this sense, updated regression estimates of form (6) may be becoming more reliable over time. This is not usually discussed as a reason for continuous updating or "sequential" estimation, for rolling regressions, or for letting estimated regression coefficients vary with the quality of information.

If this quality has in fact improved over time, changes in  $B$  do not necessarily signify changes in  $b$ . Furthermore, changes in the variance of  $Y$ —where  $Y$  could stand for GNP or the unemployment rate before and after World War I or before and after the Great Depression in recently revived debates—would not necessarily be indicative of stabilization policies or of other factors that have reduced remaining disturbances ( $dS_{ee} < 0$ ) and hence the variance of  $y$ . Rather, a decline in  $S_{YY}$  could indicate a reduction in measurement error ( $dS_{uu} < 0$ ) equally well. Thus improvements in data quality can add uncertainty about the interpretation of trends in variation if it is not known at what rate the measures are becoming better.

Adding a law of motion for the latent variable  $x$  to system (2) allows focusing on another consequence of data improvements. Changes in data quality produce time variation of the coefficients in the optimal forecast. Hence, if improvements in data quality are perceived correctly by agents, the coefficients estimated for the dynamic process they follow will change. This will happen even

when there would be no change in the dynamic forecasting equation if the true series,  $x$  and  $y$ , were known. Assuming the latent variable  $x$  follows the random walk,

$$(8) \quad x_t = x_{t-1} + z_t, \quad z_t = N(0, \sigma^2),$$

John Muth (1960) has shown that the optimal forecast involves weights on the lagged values of  $X$  that decline exponentially. Hence equation (6) is replaced by the forecasting equation,

$$(6') \quad Y_t = B(1 - c) \times (X_{t-1} + cX_{t-2} + c^2X_{t-3} \dots) + W_t,$$

where  $c$  is chosen to minimize the squared prediction error on past data. The coefficient  $c$  is lower, and hence the rate of decline of the lag weights faster, the lower the noise, due to  $S_{ee}$ ,  $S_{uu}$ , and  $S_{vv}$ , relative to the strength of the signal measured by the variance ( $\sigma^2$ ) of  $z$ . Hence if the measurement errors for  $x$  and  $y$  decline so that  $S_{uu}$  and  $S_{vv}$  fall, the response speed to lagged data would appear to rise. The reduction in the mean lag would be due not to a change in the underlying behavior, but to changes in the clarity of signals shifting the optimal balance between the risk of missing out on news and the risk of being misled by noise in recent data.

Such possible consequences of technical changes in the informational environment generally have been ignored, for instance, by those who have analyzed and interpreted the reduction in inflation of earlier this decade. In the judgment of Robert J. Gordon (1985), the rapidity of this decline surprised many who then began to look for autonomous changes in behavior rather than changes in the state of information.

## II. Temporal Compression

In addition to the possible reduction in measurement error, another important achievement in the information revolution is to make ever more data available on a timelier basis ever more frequently. Furthermore, the cost of immediate access to data increases very little with the distance from their point of release. As a result, data are now almost equally available around the

world and cheap to communicate. Under these changed circumstances, many of the signal extraction problems designed to deduce from high-frequency series available with short collection and reporting lags what may be happening to low-frequency series, or series with longer lags, now appear needlessly contrived. The survey by von Furstenberg and Jin-Ho Jeong (1988) provides numerous examples.

Signal extraction problems have been used to explain temporary disequilibria and slow adjustment to unexpected and unannounced developments. The intended lesson was that correct adjustment can be expected only if the current constellation of the data, some of which are treated as not yet observable, turns out to conform exactly to the model and the variance-covariance pattern of the residuals of past data. One would expect such a theory to imply that if more data are being reported faster and more frequently, there would be less scope for real effects to arise from errors in inferences about current data. However, any precise accounting for how changes in the actual state of information may have affected the behavior of economic time series and their model is rarely found in the signal extraction literature.

While more information can obviously solve those problems of temporary disequilibrium and slow adjustment thought to arise solely on account of deficiencies in information, it need not solve others. It can even create problems, some of which may turn out to be genuine. The reduced need for time averaging and the narrowing of time and coverage gaps in information have the effect of allowing developments to be screened almost as they happen, with less information gained longer after the fact. Indeed, certain optimization programs requiring continuous monitoring and vast computational effort are now automated and self-reprogramming, perhaps through the application of artificial intelligence. Hence distinctions between short- and long-run outcomes of impulses and disturbances that are based on the premise that "only in the long run will it all have come out" are collapsing. Furthermore, the more frequent, and therefore less averaged, the reporting of potential news and the updating of outcome distributions

like those represented by equation (1) or by more comprehensive measures of information (Maasoumi, 1988), the more fluctuating and "volatile" the level of uncertainty can be. As episodes of high news are interspersed with periods of low news intensity, the movement of high-frequency price series frequently takes on characteristics of a random walk with step-ahead distributions that are leptokurtic and inscrutable from the viewpoint of economic fundamentals.

More timely, frequent, and accurate public information that can travel at the speed of light reduces one of the objective reasons for temporary differences in subjective beliefs, that of being unavoidably differently informed at a point in time. Specialists whose definition of efficiency tends to be based on the completeness of arbitrage and the "consistency" of spot and future price series thus can undoubtedly claim that efficiency has grown in the small. On the other hand, reduced confidence in prior beliefs based on "fundamentals" can raise the volatility of outcomes. Formulas such as those derived by John Taylor (1975, p. 1017) could yield this result if there is a decline in the precision of previously held beliefs but some "cycling" or eventual reversion to the mean, as opposed to a strict random walk, in news content. Correspondingly, Sanford Grossman and Joseph Stiglitz (1976, p. 251) have suggested that added information, by increasing price variability, could raise uncertainty about, and thereby diminish, the value of one's endowments. Faster information may also lower insurability and the ability to arrive at efficient risk sharing arrangements on account of temporal compression (Jacques Drèze, 1979).

Altogether this suggests that orientation and efficiency may not be helped in all respects by the ongoing information revolution. Rather, this development can reveal as it obscures, and stabilize as it unsettles. For instance, statistical models and coefficients are likely to be made time-varying by cumulative improvements in data and the state of information, but this is rarely attended to in monitoring estimates of macroeconomic relations. Estimated relations can change and drift without any change in policy for rea-

sons quite different from those originally emphasized by Robert Lucas (1976). Losing yet one more of the things that have traditionally been taken as *pregiven* to economic analysis can be disorienting but necessary. For it is, by now, almost a contradiction in terms to treat information technology and its yield as remaining fixed long enough for macroeconomists to finish a portrait from time series under that condition.

## REFERENCES

- Aigner, Dennis J. et al., "Latent Variable Models in Econometrics," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, Vol. II, Amsterdam: North-Holland, 1984, ch. 23.
- Drèze, Jacques H., "Human Capital and Risk-Bearing," *The Geneva Papers on Risk and Insurance*, June 1979, 4, 5-22.
- Gordon, Robert J., "Understanding Inflation in the 1980s," *Brookings Papers on Economic Activity*, 1:1985, 263-99.
- Grossman, Sanford J. and Stiglitz, Joseph E., "Information and Competitive Price Systems," *American Economic Review Proceedings*, May 1976, 66, 246-53.
- Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique," in Karl Brunner and Alan Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conference Series on Public Policy, *Journal of Monetary Economics Suppl.* 1976, 1, 19-46.
- Maasoumi, Esfandiar, "Information Theory," in John Eatwell et al., eds., *The New Palgrave: A Dictionary of Economics*, Vol. 2, New York: Stockton Press, 1988.
- Muth, John F., "Optimal Properties of Exponentially Weighted Forecasts," *Journal of the American Statistical Association*, June 1960, 55, 299-306.
- Taylor, John B., "Monetary Policy During a Transition to Rational Expectations," *Journal of Political Economy*, October 1975, 83, 1009-21.
- von Furstenberg, George M. and Jeong, Jin-Ho, "Owning Up to Uncertainty in Macroeconomics," in *The Geneva Papers on Risk and Insurance*, January 1988, 13, 12-90.

# WHY IS UNEMPLOYMENT SO HIGH IN EUROPE?†

## Beyond the Natural Rate Hypothesis

By OLIVIER J. BLANCHARD AND LAWRENCE H. SUMMERS\*

In a well-known essay, Thomas Sargent (1983) treated the disinflationary policies of Britain's Thatcher government as a useful natural experiment for contrasting two alternative macroeconomic theories. On the "classical" view, considered by Sargent, disinflation, if credible, is achievable at little cost in unemployment. On the alternative "Keynesian view," even credible disinflation is likely to increase unemployment for some time, because of the inflationary momentum caused by overlapping price and wage decisions. The results from the Thatcher experiment and from similar experiments conducted in the rest of Europe are now in and the conclusion is clear: Events have proven both theories wrong. They have been proven wrong in their common presumption that, once disinflation dynamics were over, economies would be back to their previous "natural" rate of unemployment.

Consider Britain as an example. Nine years after the onset of disinflationary policies backed by two landslide election victories, and significant liberalizations in labor and product markets, and five years after the rate of wage inflation stopped declining, both theories surely would have predicted that unemployment would now have returned at least to its previous level. The reality is very different. The unemployment rate in Britain was 5 percent when disinflationary policies were commenced in 1979. Over the last four years, it has averaged 11.6 percent and stands at 10 percent today. While unemployment in

Britain is at least declining, the situation in the rest of Europe is grimmer: the OECD actually predicts an increase in the unemployment rate for OECD Europe from 11 percent in 1987 to 11.2 percent in 1988.

This paper argues that European experience of the 1980's poses a profound challenge to standard Keynesian and Classical theories of macroeconomic fluctuations. What is required is a theory of unemployment in which unemployment, far from returning to a stable equilibrium—or "natural rate"—over time, is instead strongly dependent on history. We refer to such an equilibrium as a "fragile equilibrium," to highlight the sensitive dependence of unemployment on current and past events. Developing a theory of fragile equilibrium, we argue, involves questioning and perhaps discarding traditional presumptions about the slopes of labor demand and supply curves.

Section I briefly reviews macroeconomic developments in Britain over the last ten years, highlighting the inability of standard theories to make useful contact with what has happened. Section II describes the type of fragile equilibrium theory that is necessary to understand the European experience. Section III speculates on factors that may lead to fragile equilibria, focusing on possible reasons why labor demand curves may slope upwards or labor supply curves may slope downwards. Section IV concludes by commenting on the policy implications of the results.

### I. Standard Theories Cannot Account for the European Experience

Our broad interest is in the dramatic increases in unemployment that have occurred in almost every European country. But we narrow our focus to the United Kingdom for

†*Discussants:* Lloyd Ulman, University of California-Berkeley; Robert M. Solow, MIT; Thomas E. Weisskopf, University of Michigan.

\*MIT, Cambridge, MA 02139 and NBER, and Harvard University, Cambridge, MA 02138 and NBER, respectively.



two reasons.<sup>1</sup> First, Britain is often thought to manifest the most advanced case of Eurosclerosis. Second, as emphasized by Sargent, Mrs. Thatcher's election represented a significant structural change in policy. While there can be no denying that disinflationary policies encountered credibility problems at the outset, policy surely became credible after Mrs. Thatcher's second landslide victory.

Standard theories suggest that the high current rate of unemployment could be the result of either structural factors that changed the natural rate of unemployment or cyclical factors that have temporarily driven the British unemployment rate above its normal level. Since at this late date, disinflationary policies are surely credible, Classical theories suggest that the source of high unemployment must be sought in structural factors, which have increased the equilibrium unemployment rate to its current level. But this line of reasoning runs into two central difficulties.

First, Britain has many structural problems but it is hard to see how they have gotten worse over the past decade. The last eight years have witnessed the most resolutely conservative government since World War II, three major Parliamentary Acts attacking union power, a generalized attack on the welfare state, and countless ministerial paeans to the free market. The factors stressed by those who see structural factors as a primary cause of high unemployment have all moved in the right direction. Yet more man-years of unemployment were suffered in Britain between 1979 and 1987 than in the entire 1939-79 period.

Second, the timing evidence is also difficult to square with the view that structural changes are the cause of high British unemployment. Unemployment increased by 5 percentage points between 1980 and 1982 during which time industrial production plummeted as real interest rates and the pound soared. The origins of the rise in

British unemployment in disinflationary policy are clear enough. The mystery is why unemployment has not recovered.

This leads naturally to the second standard explanation for a high unemployment rate—cyclical disturbances that have driven the unemployment rate above its natural level. Keynesian explanations for cyclical contraction are not lacking—monetary contraction at the beginning of the decade was followed by a period of fiscal austerity leading to today's high unemployment rate. The basic difficulty with this argument is that while unemployment is very high, in most other respects the British economy does not appear to be in recession.

With unemployment so far above its natural rate, inflation should be decreasing sharply. But CPI inflation in Britain stands at 4.5 percent for 1987, compared to 5 percent in 1984, and the rate of growth of nominal wages has more than kept pace with the CPI.<sup>2</sup> And, positive GNP growth since 1982—British real GNP has grown at more than a 3 percent annual rate over the last three years—has prompted widespread characterizations in Britain of the current period as the longest U.K. expansion since the war.

Labor market indicators that normally move cyclically, the rate of permanent layoffs, the overtime rate, and rate at which workers are stood off (temporarily laid off in the American terminology) all suggest a strong, not a weak, labor market.<sup>3</sup> At the beginning of 1987, for example, 34 percent of manufacturing workers worked overtime, just equal to the figure at cyclical peaks in 1974 and 1977, and significantly above the 26 percent overtime figure in 1981 and the 29 percent figure observed during the recessions in 1973 and 1976. Similarly, less than 1 percent of the work force was "stood off" during 1986. This rate is comparable to the rate observed during periods of expansion,

<sup>2</sup>More formal estimates of "non-inflationary unemployment rates" (NAIRU) confirm that in most countries, NAIRU and actual unemployment rates are close to each other.

<sup>3</sup>Summers (1988) develops the argument in this paragraph in more detail.

<sup>1</sup>For a detailed discussion of European experience, see Robert Lawrence and Charles Schultze (1987).

but less than half the rate observed during any previous recession.

The "normality" of the British labor market even in the face of very high unemployment is further evidenced by the data on workers' attitude towards the possibility of unemployment. In June 1985, 45 percent of those who were employed thought they could find a job quickly if they became unemployed compared to 40 percent in September of 1977 when the unemployment rate was only half as high. Similarly, the fraction of the employed population regarding their job as safe declined only mildly from 71 to 61 percent over the same period.

We conclude that it is difficult to account for the British experience within the standard paradigms. In many respects the British labor market appears to be in equilibrium. Yet, unemployment has doubled in the presence of structural changes that if anything should have worked to reduce it.

## II. The Need for Theories of Fragile Equilibrium

Almost any theory of the determination of wages and employment can be reduced to the intersection of two loci in wage employment space. Abusing the language somewhat, we shall refer to those loci as demand and supply. Labor demand may, for example, refer to the equilibrium locus of price and employment decisions taken by monopolistically competitive firms given the nominal wage. Labor supply may refer to the set of employment and nominal wage decisions resulting from bargaining between unions and firms given the price level.<sup>4</sup>

Put simply, standard theories of unemployment imply sharply downward-sloping labor demand curves and upward-sloping labor supply curves, leading to a sharply defined X diagram, and a unique equilibrium. They carry the implication that small shocks, small shifts in supply or demand

curves, have small effects on equilibrium unemployment. They also usually imply that this unique equilibrium is stable, and that the dynamic effects of transitory shocks on unemployment do not last very long. But, as we have just argued, these implications are precisely why standard theories cannot explain the current European experience, and why we have to look for alternatives.

A physical analogy is useful here. Consider a ball on a hilly surface. If the surface is bowl-shaped, there will be a single uniquely and sharply determined equilibrium position for the ball—at the bottom of the bowl. This is the view implicit in the natural rate hypothesis. But the European experience suggests other possibilities. If the surface contains two pronounced valleys, or is extremely flat, or contains many mild depressions, the ball's position will depend sensitively on just how the ball is shocked. We use the term "*fragile equilibria*" to refer to situations of this type—where outcomes are very sensitive to shocks, and may be history dependent.

Returning to labor supply and demand curves, the natural way to think about fragile equilibria in unemployment is to think about economic mechanisms that can give rise either to upward-sloping demand curves or downward-sloping supply curves. These schedules allow for the possibility of unstable equilibria or multiple equilibria; such equilibria have the common property of making unemployment extremely sensitive to initial conditions and to current and past shocks. Even if the equilibrium is unique and stable, equilibria will be fragile as long as the curves intersect at a narrow angle. In this case, unemployment is likely to return slowly at best to equilibrium, and depends strongly on the history of shocks.

Research on multiple equilibria (Peter Diamond, 1982) and on hysteresis (ourselves, 1986) suggest mechanisms that may generate unemployment rates that depend sensitively on the shocks an economy has experienced. As the example of the ball finding a position on a hilly surface suggests, the distinctions between these different possibilities is less important than their common implication that unemployment equilibria are likely to be fragile.

<sup>4</sup>This is, for example, the structure of the model constructed by Richard Layard and Stephen Nickell (1986) to study unemployment in the United Kingdom.

### III. Some Fragile Equilibrium Theories

Modifying the classic labor supply, labor demand determination of employment to make employment equilibria fragile requires either adducing considerations that make labor supply potentially downward sloping or labor demand upward sloping. This section suggests a number of mechanisms that can do the job.

#### A. *Downward-Sloping Labor Supply*

In our earlier papers (1986, 1987b), we built on the work of Robert Gregory (1986), in advancing the argument that there is a tendency for the equilibrium unemployment rate to track the actual unemployment rate because unions bargain only on behalf of their incumbent members. In the simplest case where union members simply set wages to insure the employment of their current members but not to permit the firm to do any hiring, the wage-setting relation and the labor demand curve will coincide allowing for a continuum of equilibria and hysteresis. Even apart from this extreme case, as long as the employed play an important role in wage setting, lower employment may breed greater not lesser aggressiveness in wage setting. As employment decreases, real wage demands by the employed workers will increase, leading to a labor supply curve which is downward sloping in wage-employment space.

There is a different way of generating very similar effects, which has been suggested by Minford. Suppose that unemployment undermines the work ethic. This may arise because of the direct effects on workers' attitudes of prolonged unemployment, because the stigma associated with unemployment declines in a high unemployment society (British libraries in the Midlands make available pamphlets with the title "Leaving School: What You Should Know About Social Security Benefits"), or because of the policy changes unemployment brings about. The Thatcher government stopped requiring unemployment insurance recipients to regularly appear at Job Centres because there was not enough room to accommodate them. In each of these cases, a reduction in em-

ployment may work to raise the wages that firms must pay to attract labor, again generating a downward-sloping labor supply schedule.

These interpretations of events in Europe would lead one to expect unemployment to be concentrated among outsiders, new entrants, or workers who have been unemployed for a long time. One would expect the employed to feel relatively secure about their jobs, overtime rather than additional hiring to be used for short-term fluctuations, and all this fits well the current situation. One would also expect the short-term unemployed to have more influence on wage bargaining than the long term unemployed and this also seems to be the case (Layard and Nickell).

A difficulty shared however by both lines of explanation is that, to the extent that they imply movements along a downward-sloping demand curve, one would have expected real wages to have increased as firms moved up their labor demand. But, by almost any measure of trend productivity growth, real wages have fallen, not risen, relative to productivity, following the disinflation that began in 1980. This suggests that they omit some important labor demand side element from the story.

#### B. *Upward-Sloping Labor Demand*

Writing from different perspectives, many authors (including Martin Weitzman, 1982; Diamond; Peter Howitt and R. Preston McAfee, 1987; and ourselves, 1987a) have recognized that increasing returns provide a natural explanation for why the labor demand curve relating the real wages of workers and the firms employment decision might slope upwards over some range. Essentially, the idea is that with increasing returns the marginal revenue product of labor increases with the level of employment. This may be because of physical increasing returns in the production process arising from fixed costs as in Weitzman, the improved matches between employees and employers suggested by Diamond's search model and explored in Howitt-McAfee, or the "fiscal increasing returns" arising from the fall in tax rates when

increases in employment raise the tax base available to finance a fixed or increasing level of government spending, as discussed in our earlier paper (1987a).

Evidence on the role of physical increasing returns is hard to adduce. The strong performance of labor productivity in U.K. manufacturing over the last five years certainly does not give strong support to this argument, though aggregate productivity growth in Europe as in the rest of the world has been slower during the 1980's when employment growth has been poor, than during previous periods when employment grew rapidly. The case for fiscal increasing returns is stronger. We demonstrated in our earlier paper that it was likely that, once tax effects were recognized, increases in employment in Europe would be associated with increases in workers' take-home pay. This suggests an upward-sloping labor demand curve in the after-tax wage-employment space.

There is a second mechanism which may generate an upward-sloping labor demand curve. Discussions of European labor markets invariably stress the adverse effects of rules that preclude firms from laying off workers. But the importance of those rules depends very much on such factors as the growth rate of demand facing firms, the quit (or natural attrition) rate, the degree of uncertainty facing firms. The lower the growth rate or the lower the quit rate, for example, the more likely a firm is to be constrained by those rules, and the greater the shadow cost they impose. This suggests a mechanism through which high unemployment may increase firm's labor costs, leading to an upward-sloping labor demand curve. As unemployment increases, workers are more reluctant to change jobs and the quit rate decreases. This increases the shadow cost of firing restrictions. If the effect of unemployment on quits, and the effect of reduced quits on the shadow cost of labor are both sufficiently strong, an upward-sloping demand curve for labor will result.

Evidence on this mechanism is also difficult to obtain. Surveys of firms suggest that the importance of firing restrictions varies across European countries, with the restric-

tions playing a minor role in the United Kingdom. Turning to direct evidence, there are no data on quits (vs. layoffs) in most European countries, including the U.K. For Italy, where data are available, the quit rate in the industrial sector has decreased from an average 14 percent during 1965-73 to an average of 6.5 percent for 1980-85. This is a substantial decrease, especially when one takes into account that, at such low levels of quit rates, the workers who quit are unlikely to be those that the firm would want to fire. The effect of the attrition rate on the shadow cost of labor is also uncertain. Recent theoretical research (Samuel Bentolila and Giuseppe Bertola, 1987) has derived the relation between the shadow cost of labor, hiring and firing costs. This research suggests that, at least for firms which have low rates of growth, decreases in the quit rate can substantially decrease average employment.

#### IV. Conclusions

We believe that understanding unemployment in Europe will require economists to dispense with the natural rate hypothesis that underlies much of both Keynesian and Classical macroeconomics. Theories of fragile equilibria are necessary to come to grips with events in Europe. We have suggested some elements that may go into the construction of these theories.

Our focus has been on unemployment equilibria. Drawing supply and demand curves with varying slopes and specifying dynamics, the reader will have no difficulty constructing examples in which equilibria exhibit a variety of stability properties. One possibility is suggested by each of the mechanisms considered above and is intriguing enough to warrant mention. Each of them suggests that the longer unemployment stays away from equilibrium, the more slowly it will eventually—if ever—return to it. Union members who lose jobs are unlikely to be disenfranchised immediately. It takes time for unemployment's stigma to diminish. Increasing returns effects will only come into play with significant lags as firms enter or exit industries, and governments adapt tax

rules. Irreversibilities in employment will weigh heavily on firms only when they expect a very long-lasting contraction in demand.

What is the policy implication of the view that European unemployment equilibria are fragile and that current high levels of unemployment are largely the legacy of past policies? Without much-needed theoretical and empirical work isolating the reasons why equilibria are fragile, it is difficult to draw firm conclusions. There are clearly two logical possibilities. These mechanisms may be such that it is more difficult to decrease than to increase unemployment. If so, there may be relatively little that macroeconomic policy can do to restore full employment. On the other hand, just as the adverse shocks of the late 1970's and early 1980's had a durable impact, policies that sharply increased the demand for labor might have the lasting effect of increasing employment. Resolving the issue will be impossible until economists are willing to move beyond the natural rate hypothesis in thinking about European unemployment.

#### REFERENCES

- Bentolila, Samuel and Bertola, Giuseppe, "Firing Costs and Labor Demand in Europe: How Bad is Eurosclerosis?," mimeo., MIT 1987.
- Blanchard, Olivier and Summers, Lawrence, "Hysteresis and the European Unemployment Problem," *NBER Macroeconomics Annual* 1986, Cambridge: MIT Press, 1986, 15-78.
- \_\_\_\_\_, and \_\_\_\_\_, (1987a) "Fiscal Increasing Returns, Real Wages, and European Unemployment," *European Economic Review*, April 1987, 31, 543-66.
- \_\_\_\_\_, and \_\_\_\_\_, (1987b) "Hysteresis in Unemployment," *European Economic Review*, February/March 1987, 31, 288-95.
- Diamond, Peter, "Aggregate Demand in Search Equilibrium," *Journal of Political Economy*, October 1982, 90, 881-94.
- Gregory, Robert, "Wage Policy and Unemployment in Australia," *Economica*, February 1986, 53, S53-74.
- Howitt, Peter, and McAfee, R. Preston, "Costly Search and Recruiting," *International Economic Review*, February 1987, 33, 89-107.
- Lawrence, Robert and Schultze, Charles, *Barriers to European Growth: A Transatlantic View*, Washington: The Brookings Institution, 1987.
- Layard, Richard and Nickell, Stephen, "Unemployment in Britain," *Economica*, February 1986, 53, S121-71.
- Sargent, Thomas, "Stopping Moderate Inflation: The Methods of Poincare and Thatcher" in R. Dornbusch and M. Simonsen, eds., *Inflation, Debt and Indexation*, Cambridge: MIT Press, 1983, 54-99.
- Summers, Lawrence, "Hysteresis and British Unemployment," Employment Institute Pamphlet, London, 1988.
- Weitzman, Martin, "Increasing Returns and the Foundations of Unemployment Theory," *Economic Journal*, December 1982, 92, 787-804.

# Is European Unemployment Classical or Keynesian?

By ROBERT M. COEN AND BERT G. HICKMAN\*

European unemployment rose sharply beginning in the mid-1970's and remains very high today, yet policymakers seem reluctant to pursue fiscal and monetary actions that would greatly stimulate aggregate demand. Traditional Keynesian policies may be impeded by a belief that rising unemployment is not so much Keynesian, arising from shortfalls in aggregate demand, as it is "classical," originating from a failure of real wages to adjust to changing market conditions, particularly reduced rates of productivity growth.

To determine the extent of classical vs. Keynesian unemployment, we apply a theoretical approach developed by Hickman (1987) to econometric models of labor supply and demand in selected European countries and the United States. The models are used to determine the time paths of the natural unemployment rate, potential output, and the full-employment real wage that clears the labor market at the natural rate of unemployment. The computed potential paths are benchmarks for measuring shortfalls of aggregate demand and excesses of actual over full-employment real wages. Effects of eliminating wage gaps are studied in counterfactual simulations, which provide information for decomposing unemployment into its natural, Keynesian, and classical components.

## I. Concepts of Classical and Keynesian Unemployment

In the standard fix-price model of price-taking competitive firms, Keynesian and classical unemployment are separate states according to whether notional product supply exceeds or falls short of market de-

mand at the prevailing wage and price configuration, so that labor demand is either output constrained and determined by the inverted production function (Keynesian unemployment), or firms are on their notional product supply and labor demand functions but the real wage exceeds the Walrasian full-employment level (classical unemployment). Thus labor demand is independent of the real wage in the Keynesian state and depends only on the real wage in the classical state.

Our conceptual framework assumes instead that imperfectly competitive firms price according to a markup rule and choose their inputs of labor and capital to minimize the cost of producing the output they expect to sell at the price they have set. The labor demand function is always conditional on both output and the real wage, and classical and Keynesian unemployment may therefore coexist. Keynesian unemployment exists when the economy is operating on an isoquant below that representing potential output, whereas classical unemployment occurs when the wage-rental ratio exceeds its full-employment level either at or below potential output.

It is important to stress that the concepts refer to employment states rather than to the shocks which induced them. A wage shock, for example, may not only increase the real wage above the full-employment level, but also depress aggregate demand because of negative Keynes or Pigou effects. Conversely, a deflationary monetary or fiscal shock may depress the real wage as well as affecting aggregate demand and output. Thus either type of shock may induce both types of unemployment.

## II. Empirical Models of Labor Supply and Demand

The specifications and sample periods of our models differ somewhat from country to

\*Departments of Economics, Northwestern University, Evanston, IL 60201, and Stanford University, Stanford, CA 94305. The support of the Austrian National Bank is gratefully acknowledged.

country (see our 1987 article for details), but their main features can be briefly summarized.

On the supply side of the labor market, the labor force is disaggregated by age and sex, usually into 16 groups. Aggregate labor force is given by

$$(1) \quad L = \sum_i N_i LP_i [(1 - \tau)(W/PC)],$$

$$(E/N), AH, t, Z],$$

where  $N_i$  is the population in group  $i$ ,  $LP_i$  is the estimated participation equation for group  $i$ ; and  $\tau$  is the personal tax rate,  $W$  the nominal hourly wage,  $PC$  the consumption deflator,  $E$  aggregate employment,  $N$  aggregate population,  $AH$  annual hours per worker,  $t$  a time trend, and  $Z$  a vector of other exogenous variables. The determinants of participation of main interest here are the after-tax real wage, measuring the opportunity cost of leisure, and the aggregate employment ratio which, as an indicator of the probability of finding work, captures discouraged-worker effects.

Workers' desired hours are also assumed to depend on the after-tax consumption wage, but cyclical variations in labor demand, proxied by the unemployment rate,  $U$ , may affect actual hours. Hours may also vary inversely with the proportions of women and teens in the labor force,  $LW$  and  $LT$ , since these groups are more likely to engage in part-time work. Thus, the average hours equation takes the general form:

$$(2) \quad AH = AH[(1 - \tau)(W/PC),$$

$$U, LW, LT].$$

The full-employment labor supply is the quantity of man-hours that would be supplied at the natural rate of unemployment. Our measure of the natural rate is based on Michael Wachter's (1976) work and allows for the effects of demographic changes on structural unemployment, but it is not defined or estimated from a Phillips curve as a NAIRU. We distrust NAIRUs as policy guides because empirical estimates of their

magnitude are subject to wide variation according to differing attempts to adjust for shifts in the Phillips curve in response to supply shocks and imprecise coefficient estimates (see David Coe and Francesco Gagliardi, 1985).

The aggregate natural rate  $UF$  is a weighted average of the age-sex specific natural rates,  $UF_i$ :

$$(3) \quad UF = \sum_i UF_i (LF_i/LF),$$

where  $LF_i/LF$  is the fraction of the full-employment labor force in the  $i$ th age-sex group. The  $UF_i$  are computed from regressions of each group's actual unemployment rate on the unemployment rate of prime-age males and on the fraction of the population in group  $i$ . The population ratio is intended to capture structural or frictional variations in the age-sex specific rates. On the assumptions that the prime-age male rate is largely unaffected by structural changes in the labor market and that its natural level is constant over time at a value observed in a period that is generally acknowledged to be one of high employment, the natural rate of prime-age males is substituted into the regressions to purge the observed unemployment rates of their cyclical components.

We shall denote full-employment levels of variables by the suffix  $F$ . Upon substituting  $E = EF = (1 - UF)LF$  in (1) and  $U = UF$  in (2) and affixing  $F$ 's to the other variables, the system can be solved for  $LF$ ,  $EF$ ,  $UF$ , and full-employment man-hours  $MHF = (AHF)(EF)$ , conditional on the real wage and exogenous variables.

The demand for man-hours is derived on the assumption that firms choose capital and labor inputs to minimize the expected costs of producing expected output, subject to a Cobb-Douglas production function with constant returns to scale. Thus, desired man-hour input depends on expected output, relative factor prices, and the level of total factor productivity. Because of adjustment costs, firms are assumed to close only a fraction of the gap between desired and actual man-hours each period. Thus, the dis-

equilibrium demand function is

$$(4) \quad MH = k[e^{-\gamma}(WE/QE)^{-\alpha}XE]^{\lambda} \\ \times MH_{-1}^{1-\lambda},$$

where  $k$  is a constant,  $\gamma$  is the rate of Hicks-neutral technical progress,  $WE$  is expected wage,  $QE$  is expected implicit rental price of capital,  $XE$  is expected output,  $\alpha$  is capital elasticity in the production function, and  $\lambda$  is speed of adjustment. The implicit rental price of capital is defined by  $QE = PIE(r + d)T$ , where  $PIE$  is expected price of new capital goods,  $r$  the discount rate of firms,  $d$  the depreciation rate, and  $T$  measures effects of the tax treatment of investment expenditures. Our tests of alternative measures of  $r$  indicate that the preferred assumption is that  $r$  is constant (6 percent per annum is used here). Expected values of wages, prices, and output are generally determined by predictions from autoregressions.

Substituting  $MHF$  on both sides of (4), and assuming that expectations would be realized along a full-employment growth path, we invert the labor demand equation to obtain an expression for potential output,  $XP$ , defined as the level of output that would clear the labor market at the natural rate of unemployment:

$$(5) \quad XP = k^{-1}e^{\gamma}(W/Q)^{\alpha}MHF^{1/\lambda} \\ \times MHF_{-1}^{-(1-\lambda)/\lambda}.$$

$XP$  is conditional on the real investment wage and on the real consumption wage, which is a determinant of  $MHF$ . To close the system, we assume that along the natural growth path, real wages (both  $W/PC$  and  $W/PI$ ) grow at the same rate as potential labor productivity,  $XP/MHF$ .

A key determinant of the growth of potential productivity and therefore of the full-employment wage is the rate of technical progress which, according to our estimates of labor demand functions, declined for all countries following the first oil shock in 1973-74. Additionally, the oil shock was associated with sharp increases in prices of

TABLE 1—EXCESS OF ACTUAL OVER  
FULL-EMPLOYMENT REAL INVESTMENT WAGE

Country	1973	1974-79	1980-83	1984
Austria	-0.6	9.4	7.3	3.0
Germany	8.0	12.8	3.4	-0.2
U.K.	9.2	10.7	12.2	15.9
U.S.	6.2	8.8	11.6	11.2

Note: Shown in percent.

imports and investment goods relative to money wages in the United Kingdom and United States, producing abrupt and lasting declines in ratios of real wages to productivity. Accordingly, we introduce exogenous downward adjustments in full-employment real wages in these countries at the time of the shock. Exogenous upward adjustments are made for the United States in 1981-84 to account for the extraordinary increases in the real import price and in the real investment wage relative to productivity growth accompanying the sharp appreciation of the dollar. The estimated investment wage gaps during 1973-84 are shown in Table 1.

### III. Decomposing Unemployment

To illustrate our theoretical decomposition of unemployment, let us abstract from variations in average hours, which are taken into account in the empirical estimates, and measure labor demand and supply in numbers of workers. A state of excess unemployment is depicted in Figure 1 (reproduced from Hickman, 1987, p. 1546).  $LDP$  is the labor demand schedule at potential output. The full-employment labor force is  $LF$  (drawn as wage-inelastic for simplicity), and the full-employment wage  $WRF$  would clear the labor market at the natural rate of unemployment  $(LF - EF)/LF$ .

The actual labor demand function whose position depends on current output and lagged employment is  $LD$ , and  $E$  persons are employed at the current real wage  $WR$ . There is a shortfall of  $(EF - E)$  between actual and full-employment employment. The actual or measured labor force  $L$  depends on the level of  $E$ , owing to the discouraged-worker effect, which generates hid-



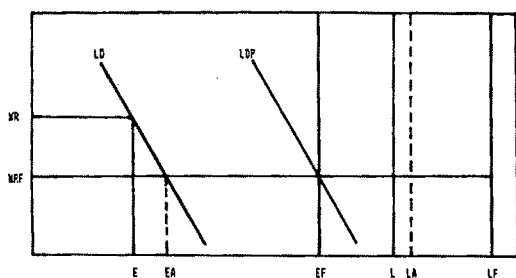


FIGURE 1

den unemployment of  $(LF - L)$ . Actual or measured unemployment is  $(L - E)$ , and the measured unemployment rate is  $(L - E)/L$ . The larger is the elasticity of labor force with respect to employment, the larger will be the increase in hidden unemployment and the smaller will be the increase in measured unemployment for a given decline in labor demand.

The magnitude of discouraged-worker effects varies among countries, as well as over time for a given country as the composition of the workforce changes. To facilitate comparisons of unemployment, we adopt for our analysis the "adjusted unemployment rate,"  $(LF - E)/LF$ , obtained by adding hidden unemployment to the numerator and denominator of the actual rate.

As a counterfactual experiment, suppose that the market wage were reduced to  $WRF$  with aggregate demand and output unchanged. Employment would then equal  $EA$  instead of  $E$ , so that of the original employment shortfall of  $(EF - E)$ ,  $(EA - E)$  is the classical component, attributable to the wage gap  $(WR - WRF)$ , and  $(EF - EA)$  is the Keynesian element, assignable to deficient product demand. The adjusted unemployment rate may then be factored into its natural, Keynesian, and classical components:

$$(6) \quad (LF - E)/LF = (LF - EF)$$

$$/LF + (EF - EA)/LF + (EA - E)/LF.$$

Note that workers would enter the labor force as employment rose in response to the wage cut, so that unemployment would fall

only to  $(LA - EA)$ , instead of  $(L - EA)$ , owing to the induced reduction of hidden unemployment by  $(LA - L)$ .

To calculate  $EA$ , the labor demand and supply functions and associated identities were solved simultaneously, setting real wages at their full-employment levels, and using the actual values of output and other variables.

#### IV. Actual, Natural, and Adjusted Unemployment Rates

Table 2 shows the uptrend of actual unemployment since the first energy shock in 1973. The actual and natural rates diverged substantially in Germany, the United Kingdom and the United States after 1973 and in Austria after 1980. The adjustment for hidden unemployment adds about 1 percentage point to the average unemployment rate for Germany, the United Kingdom, and the United States in 1974-83, and up to 2 points in 1984. About one-fifth of total unemployment was hidden in these countries. In Austria, hidden unemployment was negative until 1981, but by 1984 it was adding 2 percentage points to actual unemployment and accounting for nearly two-fifths of the total unemployment rate.

TABLE 2—ACTUAL AND ESTIMATED UNEMPLOYMENT RATES

Rate	1973	1974-79	1980-83	1984
<b>Austria</b>				
Natural	1.6	1.7	1.7	1.8
Actual	1.0	1.6	2.7	3.9
Adjusted	-0.1	-0.2	2.7	6.0
<b>Germany</b>				
Natural	1.0	1.1	1.2	1.3
Actual	1.0	3.6	5.8	8.4
Adjusted	1.8	4.6	6.8	9.7
<b>United Kingdom</b>				
Natural	1.6	1.8	2.0	1.8
Actual	2.2	4.2	9.2	11.3
Adjusted	3.1	5.0	10.7	13.5
<b>United States</b>				
Natural	6.0	6.0	5.7	5.2
Actual	4.9	6.8	8.5	7.5
Adjusted	6.7	8.0	9.6	9.1

Note: Shown in percent.

TABLE 3—NATURAL, DEMAND AND WAGE SHARES OF ADJUSTED UNEMPLOYMENT

Share	1973	1974-79	1980-83	1984
<b>Austria</b>				
Natural	—	—	63	29
Demand	—	—	30	57
Wage	—	—	7	14
<b>Germany</b>				
Natural	56	25	18	13
Demand	7	47	76	83
Wage	37	28	6	4
<b>United Kingdom</b>				
Natural	52	36	19	13
Demand	—14	49	72	62
Wage	63	15	9	25
<b>United States</b>				
Natural	90	75	59	57
Demand	8	19	39	38
Wage	2	6	2	5

Note: Shown in percent.

#### V. Classical and Keynesian Unemployment.

Table 3 presents the decomposition of adjusted unemployment rates into their natural, demand, and wage components, as in equation (6). Natural unemployment accounted for a quarter of the adjusted unemployment in Germany during 1974-79. The remaining employment shortfall of 75 percent was divided between a Keynesian share of 47 percent and a classical component of 28 percent. The figures for the United Kingdom are roughly comparable, whereas excess unemployment over the natural rate was much smaller in the United States. Austria kept unemployment at about the natural rate until 1980 (Table 2) by offsetting the real wage gaps which emerged after 1973 with positive demand gaps.

Adjusted unemployment increased markedly in the European countries during the 1980's while natural rates were virtually unchanged (Table 2); consequently, the shares of natural in adjusted unemployment declined. Our estimates assign most of the increase in excess unemployment to the demand components. Adjusted unemployment also increased in the United States, again with only a minor classical contribution, but with a larger natural component.

The magnitude of the wage component of the employment shortfall,  $(EA - E)/(EF - E)$ , depends on the wage elasticity of labor demand as well as on the size of the wage gap. The estimated short-run wage elasticities range only between  $-.16$  for the United States and  $-.22$  for the United Kingdom. Substantial classical unemployment may nonetheless result from large wage gaps, as seen in Table 3.

An additional complication is introduced by the endogenous response of average working hours to wage changes. In all four countries, the partial wage elasticity is negative, indicating a dominant income effect, so that the favorable employment effect of a real wage reduction on man-hour demand is partly nullified by an induced increase of average hours.

#### VI. Utilization Rates

Our concept of potential output is a notional path assuming optimal trajectories for capital and labor under continuous full-employment equilibrium. Hence, it is not necessarily an indicator of the demand expansion required to eliminate Keynesian unemployment in a particular year. Full-employment output, defined as the output required to employ the full-employment labor supply at the full-employment real wage, conditional on the *actual* capital and labor inputs in the preceding year, is such an indicator.

On either criterion, our empirical measures suggest substantial slack for demand management policies to combat Keynesian

TABLE 4—UTILIZATION RATES

Country	1973	1974-79	1980-83	1984
<b>Potential Output</b>				
Austria	102	103	101	98
Germany	99	99	95	92
U.K.	102	96	87	85
U.S.	102	100	96	95
<b>Full-Employment Output</b>				
Austria	102	103	100	95
Germany	99	96	92	89
U.K.	101	95	86	83
U.S.	101	99	95	92

Note: Shown in percent.

unemployment in the 1980's (Table 4), particularly in Germany and the United Kingdom; and despite considerable growth of actual output in the United States in 1983-84, utilization rates in the United States remain low by comparison with the 1970's.

### VII. Some Policy Implications

We have already noted that our concepts of Keynesian and classical unemployment refer to economic states rather than to the nature of the underlying shocks. Keynesian unemployment, for example, may result from supply shocks, owing directly to real balance or wealth effects and indirectly to non-accommodating monetary policies. By the same token, the existence of substantial Keynesian unemployment does not necessarily imply that expansionary policies can eliminate the shortfall without inducing unacceptable additional classical unemployment or generating inflationary pressures. But the extent of recent output gaps in the European countries, together with the observation that real wages may not be much altered by demand management, lead us to conclude that there is substantial room for demand expansion to reduce unemployment

and a correspondingly large degree of excess capacity to restrain inflationary pressures. However, a full assessment of the inflationary consequences of a managed approach to full employment requires a complete model endogenizing the response of nominal wages and prices as well as aggregate demand and real wages to fiscal and monetary policies; hence, a complete policy prescription is beyond the scope of this paper.

### REFERENCES

- Coe, D. T. and Gagliardi, F., "Nominal Wage Determination in Ten OECD Economies," OECD Economics and Statistics Working Paper 19, Paris, March 1985.
- Coen, R. M. and Hickman, B. G., "Keynesian and Classical Unemployment in Four Countries," *Brookings Papers on Economic Activity*, 1:1987, 123-93.
- Hickman, B. G., "Real Wages, Aggregate Demand, and Unemployment," *European Economic Review*, December 1987, 31, 1531-60.
- Wachter, M. L., "The Changing Cyclical Responsiveness of Wage Inflation," *Brookings Papers on Economic Activity*, 1:1976, 115-59.

# West European Unemployment: Corporatism and Structural Change

By ANDREW GLYN AND BOB ROWTHORN\*

The current unemployment crisis of the advanced capitalist economies is primarily a European phenomenon. Between 1973 and 1986 the average unemployment rate in OECD Europe rose, from 3 to 11 percent, compared to a rise of 2 percentage points in both the United States and Japan. Nor have the European economies been providing jobs for people previously outside the measured labor force. The ratio of employment to population of working age (the "employment rate") declined even faster than unemployment rose, since average participation rates fell. By contrast, the U.S. employment rate rose faster after 1973 than before, and in Japan the employment rate stopped declining.

A simple decomposition of the growth of the labor force into population and participation rate changes (Table 1) suggests some mitigating circumstances. Europe suffered a noticeable acceleration in the rate of growth of the population of working age after 1973, while in both Japan and the United States there was a sharp decline in the rate of population growth. In the case of Japan, this decline was so great that unemployment remained low despite a marked slowdown in the growth rate of employment. In the United States, the slowdown in employment growth was somewhat less than in Europe. But the proximate explanation is hardly to the discredit of the European economies. As Table 1 shows, the ability of the United States to keep employment expanding reflects the fact that U.S. productivity

growth slowed down virtually in step with output growth after 1973. In Europe, by contrast, the slowdown in productivity growth was less severe than in output growth leaving employment to take more of the strain. Despite all the talk of "Euroscelerosis" the decline in productivity growth was *proportionately* less than in the United States (and left the average growth rate three times the U.S.'s minimal 0.6 percent). Unfortunately, in the context of stagnant output even partial success in maintaining productivity growth was at the expense of jobs.

## I. Diversity of European Experience

The above considerations help to place European unemployment in perspective. It is just as important, however, to appreciate that there has been a spectacular diversity of unemployment experience within Europe itself. By 1985, unemployment ranged from under 3 percent in Switzerland, Norway and Sweden, to 3–5 percent in Austria and Finland, 7–10 percent in Denmark and Germany, 10–13 percent in Italy, France U.K., Belgium and the Netherlands and finally 17.5 percent in Ireland and 22 percent in Spain.<sup>1</sup> Explicit analysis of this diversity has received surprisingly little attention from economists.

Broad indicators of economic performance are only weakly correlated with the pattern of unemployment increases since 1973. Equation 1 in Table 2 attempts to account for unemployment increases since 1973 across 19 OECD countries (14 European economies plus the United States, Canada

\*Corpus Christi College, Oxford OX1 4JF and Faculty of Economics and Politics, Sidgwick Avenue, Cambridge CB3 9DD, respectively. We are grateful to the World Institute of Development Economic Research, Helsinki, for financial support and to other members of its research program on Global Macro Economic Policies for comments.

<sup>1</sup>Excluded are tiny Luxembourg, mainly agrarian Greece, and Portugal because reliable intertemporal comparisons are made impossible by the 1974 revolution.

TABLE 1

Average Percent Growth Rates	Europe	U.S.	Japan
<b>POPULATION</b>			
1960-73	0.7	1.7	1.7
1973-85	1.0	1.4	0.9
Change	0.3	-0.3	-0.8
<b>PARTICIPATION RATE</b>			
1960-73	-0.3	0.3	-0.4
1973-85	-0.2	0.7	0.1
Change	0.1	0.4	0.5
<b>LABOR FORCE</b>			
1960-73	0.4	2.0	1.3
1973-85	0.8	2.1	1.0
Change	0.4	0.1	-0.3
<b>GDP</b>			
1960-73	4.7	3.9	9.6
1973-85	1.9	2.5	3.8
Change	-2.8	-1.4	-5.8
<b>PRODUCTIVITY</b>			
1960-73	4.3	1.9	8.3
1973-85	1.8	0.6	3.0
Change	-2.5	-1.3	-5.3
<b>EMPLOYMENT</b>			
1960-73	0.4	2.0	1.3
1973-85	0.1	1.9	0.8
Change	-0.3	-0.1	-0.5
<b>UNEMPLOYMENT</b>			
1960-73 <sup>a</sup>	0.0	0.0	0.0
1973-85 <sup>a</sup>	0.6	0.2	0.1
Change	0.6	0.2	0.1

Source: OECD Labour Force Statistics, Historical Statistics.

<sup>a</sup>These statistics refer to the annual increase in the percentage unemployment rate (i.e., first difference); all other statistics refer to proportionate growth rates.

Japan, New Zealand, and Australia). The explanatory variables are *changes* in the growth rate of population of working age, GDP, and product wages.<sup>2</sup>

The influence of population growth on unemployment after 1973 has already been mentioned in the comparison between Europe and Japan. In fact, this variable is easily the most significant in equation 1 and

<sup>2</sup>Changes in growth rates (as compared to the period 1960-73) rather than the growth rates themselves are used since changes in performance seem more likely to account for changes in unemployment after 1973, bearing in mind that during the earlier period, unemployment rates were mostly fairly constant. Our statistical tests, reported fully in our forthcoming article, confirmed this.

TABLE 2—UNEMPLOYMENT EQUATIONS:  
19 OECD COUNTRIES

	(1)	(2)	(3)
Time period	1973-85	1973-85	1979-85
Constant	0.305	0.018	-0.241
GDP	-0.157		
	(2.2)		
PROD WAGES	0.074		
	(1.0)		
POPULATION	0.541	0.428	0.258
	(3.3)	(2.9)	(1.6)
INDUSTRIAL EMPLOYMENT		-0.230	-0.349
		(3.8)	(6.3)
SERVICES EMPLOYMENT		-0.112	0.032
		(1.3)	(0.3)
R <sup>2</sup>	0.353	0.726	0.762

Notes: Dependent variable is the average annual increment in the unemployment rate (i.e., first difference); Independent variables are *changes* in percent per annum growth rates (as compared to 1960-73); *t*-values are in parentheses; Switzerland is omitted from equations 2 and 3 (see text).

represents a factor generally ignored in discussions of rising unemployment.<sup>3</sup>

The change in the growth rate of GDP is just about significant in accounting for increased unemployment since 1973. The correlation is quite weak, however, showing that a wide range of unemployment outcomes is consistent with a given output performance.

The failure of real wage growth to respond flexibly to productivity slowdown and unfavorable terms of trade is a popular explanation for rising unemployment. However, the increased unemployment over the period 1973-85 is only weakly correlated with variations in product wages growth. Product wages were statistically significant for the period 1973-79 (not shown here), but not thereafter. It should be noted that the conception of the role of product wages that generally inspires the use of this pre-tax variable (i.e., that of ensuring the profitability of marginal production in the market sector) is

<sup>3</sup>This is probably because attention is usually concentrated on the labor force, the growth of which, for a given growth of population, tends to be cut by unemployment as participation rates are reduced.

rather narrow. A fuller treatment should take into account in particular the role of taxation on labor as a means of providing the resources for state sector employment (as we illustrate in the final section).

It would doubtless be possible to improve on the degree "explanation" of unemployment diversity by more sophisticated modeling of the macroeconomic context. Yet this simple analysis provides important information. While confirming that a neglected factor (changes in population growth) has played a substantial role, it also suggests that some fashionable variables have been of secondary importance in explaining why some countries have been more successful than others in containing unemployment.<sup>4</sup> Within a given pattern of macro performance, a wide range of unemployment outcomes are possible depending upon the social and political institutions of the country concerned. Before discussing this further, we need to examine another neglected issue—the role of structural change.

## II. Employment Patterns

Labor can in principle move between sectors of the economy. It would therefore be natural to expect that unemployment would be more closely correlated with total employment growth (or its rate of change) rather than with employment in any particular sector. This is not so. Between 1973 and 1985, changes in the growth rate of industrial employment explain a considerably higher proportion of the variance in unemployment than do changes in the growth rate of total employment (47 as compared to 35 percent).

<sup>4</sup>The degree of explanation in equation 1 may appear to be very low in comparison with the unemployment equations estimated by M. Bruno (1986), for example. It should be noted, however, that his results are for pooled time-series and cross-section data and it seems likely that a disproportionate amount of the variance being explained is the time-series component. The evolution of the pattern of unemployment over time *within* countries may well be correlated with variables with little or no explanatory power in explaining unemployment differences *between* countries (such as the growth of world trade).

Equation 2 in Table 2 includes as explanatory variables changes in the growth rates of population and service employment as well as industrial employment. The coefficients indicate that every 1 percent per year speed-up in population growth contributed 0.43 percent per year to unemployment, while a 1 percent per year slowdown in the provision of industrial jobs raised unemployment by 0.23 percent per year (about twice the statistically insignificant effect for services employment). Since industrial employment was typically about one-third of the total, this implies that most of any slowdown in industrial employment was reflected in rising unemployment. As can be seen from equation 3 in Table 2, the importance of industrial employment is even greater in the subperiod 1979–85. This variable alone accounts for 73 percent of the variance in unemployment performance during this subperiod.

The greater importance of industrial jobs in determining measured unemployment may be explained as follows. Among industrial workers, the vast majority are full time; their skills are often specific to industrial work; moreover, industrial employment is often geographically concentrated in particular areas. When there is a major decline in industrial employment, this cannot be achieved through natural wastage, but only through wholesale redundancies in which large numbers of industrial workers are laid off. As a result, the local labor market in the industrial areas may be flooded with relatively immobile middle-aged workers, many lacking the skills required for immediate redeployment elsewhere in the economy. Even when the decline in industrial employment is achieved by natural wastage, the result is a drying up of job opportunities for young people in the area.

A rapid decline in industrial employment is unlikely to be offset on anything like an adequate scale by rising service employment. Regions that experience the greatest decline in industrial employment will typically experience the slowest growth in service employment, so unemployment in these regions will rise sharply. Moreover, even where the growth of service employment is quite fast,

the kind of labor required may be very different from that displaced from the industrial sector. Not only are the required skills often different, but many of the new service jobs created nowadays are part time and do not provide an adequate replacement for full-time industrial employment. As a result, they are frequently occupied by married women drawn back into the labor force. This explains why the growth of service employment has had only limited impact on the unemployment created by loss of industrial jobs in recent years.

The rise in unemployment, especially after 1979, has substantially the character of an industrial crisis. A number of countries (Spain, U.K., Belgium, Ireland, Netherlands, France) have lost industrial jobs at a very rapid rate (2.4 to 3.8 percent per year), and unemployment has increased sharply even in those countries (U.K. and Netherlands) where services employment continued to expand as fast or faster than before 1973.

Since around three-quarters of industrial jobs are occupied by men, it would seem that male unemployment would be more affected by a reduction in industrial employment than would female unemployment. Yet reestimation of equation 2 in Table 2 separately for men and women indicates that industrial employment is just as important in accounting for female unemployment as male. The reason would seem to be as follows. A substantial part of the female labor force consists of women who require full-time work and who will not readily accept part-time jobs. Most industrial employment is full time, and so a reduction in this type of employment directly reduces the number of full-time jobs available for women. Moreover if the decline in industrial employment is geographically concentrated, it will have a knock-on effect on some types of service employment (such as distribution) through its effect on local incomes, thereby reducing still further the amount of full-time employment available for women. Finally some of those full-time service jobs that are created may go to married women not previously in the labor force, rather than to displaced industrial workers. These factors help to explain why a rapid decline in industrial em-

ployment may lead to a considerable rise in measured unemployment amongst women, even though in aggregate it may be accompanied by a fairly large increase in total female employment. The increase in *measured* unemployment refers mainly to women workers affected by the decline in industrial jobs, while the simultaneous creation of new service jobs may provide employment for women not previously classified as in the labor force.

We now explain briefly how our analysis relates to the NAIRU approach which has increasingly dominated academic writing on unemployment (see Rowthorn, 1977; C. R. Bean, P. R. G. Layard, and S. J. Nickell, 1986). A frequent criticism of empirical work in this tradition is that there are enormous variations in the NAIRU through time and between countries that are left more or less unexplained. Our results may help to remedy this. For example, a rapid decline in industrial employment may create a pool of long-term unemployed that is not easily removed when jobs are created elsewhere. The members of this pool may be isolated from the central core of the labor market and have little effect on wage bargaining. In the NAIRU literature, this would be classified as an example of hysteresis, whereby a rise in the actual rate of unemployment increases the "equilibrium" rate. Thus, our account is broadly consistent with the NAIRU approach. Our particular contribution is to emphasize the role of structural change. Most of the NAIRU literature is highly aggregative in character and relies almost exclusively on purely macroeconomic variables. In contrast, we stress the role of structural change as a cause of unemployment, and also the extent to which the NAIRU can be influenced by policies for controlling the pace and form of such change.

### III. Corporatism and the "Star Performers"

From the above discussion it is clear that labor market performance cannot be assessed only on the basis of the official unemployment rate—the level and growth rate of employment, both absolutely and in relation to population of working age should also be

analyzed. Among the European economies, Sweden and Norway stand out. Not only have they maintained very low unemployment rates, but they have also recorded the largest increases in employment rates amongst the OECD countries (8.2 and 12.9 percent, respectively, between 1973 and 1985) and the biggest increases in participation rates from already very high levels. The other economy with very low unemployment, Switzerland, has a poor employment record and has kept people off the unemployment count by repatriating foreign workers, then by holding down the participation of women.<sup>5</sup> Austria has been relatively successful in holding down unemployment, but this has been achieved partially through Swiss-type policies where the jobs of male nationals have been protected at the expense of foreigners and women. Finland's performance has been much more impressive, especially since 1979, but space dictates confining our remarks mainly to the more outstanding cases of Sweden and Norway.

These two countries exhibited quite contrasting patterns. Norway achieved exceptional growth of *GDP* by European standards as North Sea oil production boosted industrial (but not manufacturing) output. Between 1977, when oil and gas production began to build up, and 1985, industrial production rose by 44 percent. Yet, during this period, there was no increase whatsoever in the take-home pay of most workers. Instead, the huge oil and gas revenues were deployed towards social objectives including raising farm incomes by some 50 percent to stem the outflow of population from the countryside, expanding employment in the public sector, especially of women, and subsidizing the geographically dispersed manufacturing sector (where employment declined by only 6 percent). This is in stark contrast to U.K. policy under Thatcher, whose government has positively encouraged firms to lay off

workers while at the same time reducing government employment.

The Swedish case is even more remarkable, because full employment was maintained without the benefits of oil rents. Sweden's *GDP* growth was average and growth of industrial output was slow. But, despite this, the reduction of industrial jobs was only moderate as a massive program of job protection was undertaken following the 1973 oil shock. The idea was to preserve jobs in sectors such as steel and shipbuilding while retraining workers and developing new industries. This policy was successful, as even former critics now admit (for example, OECD, 1985). All this was achieved without the wholesale shake-out which occurred in many other European economies faced with similar difficulties, such as Belgium and the United Kingdom. The crucial factor behind the very rapid expansion of the services sector was government employment that rose by well over one-third. To finance this program, Swedish workers accepted a cut in take-home pay of more than 10 percent between 1977 and 1984, mainly through increased direct taxation (see OECD, 1984).

In both Sweden and Norway, durable compromises between employers and workers have allowed full employment to be maintained as a main objective. Centrally organized workers and employers represent broad social, rather than sectional, interests. In particular the strong centralized trade union movement allows the working class to act as a class and impose full-employment policies as a price of social peace. In both countries an exceptional sense of social solidarity (and appreciation of the benefits from the welfare state) has allowed resources to be diverted to expanding state services.

The political basis of these developments has been extensively examined by writers on corporatism (see J. E. Goldthorpe, 1984). As interpreted by many economists (for example, Bruno and J. Sachs, 1985), corporatism has a rather narrow function. In line with traditional stress on the pre-tax wage as determining unemployment, the role of corporatist institutions most emphasized is the organization of restraint in wage bargaining in order to ensure the profitability of mar-

<sup>5</sup>Between 1973 and 1977, total employment fell by 280,000 and the number of foreign workers by 251,000. Switzerland was the only country other than Austria where female participation rates have fallen since 1973.



ginal production in the private sector. The examples of Sweden and Norway, however, demand a much broader interpretation. The working class has exerted "post-tax" wage restraints in order to finance the expansion of employment in the public services and programs to temper the rate of, and ease the burdens of, structural change (slowing down the decline of old sectors and, in the case of Sweden, financing retraining programs).

Over the long term, of course, it is not sufficient merely to protect existing employment and create new jobs in the public services. For the corporatist compromise to survive, the underlying problems of the economy must be tackled. An acceptable rate of growth must be achieved and fundamental weaknesses in economic structure eliminated. If this is not done, the institutions of social solidarity will come under increasing strain and eventually disintegrate. Some corporatist countries, notably Sweden and Finland, have managed to restructure their economies successfully and the social compromise is likely to endure. In Norway the future is not yet clear, as the country is only now having to confront the problems created by falling oil prices. Finally, there is Austria, where the corporatist compromise is in severe difficulties because of the country's prolonged failure to accept the need for structural change and to organize it in a humane manner. After years of hesitation, changes are now being rushed through, caus-

ing mass redundancies and large-scale unemployment. Austria illustrates the conservative dangers of corporatism—Finland and Sweden illustrate its potential for dynamism.

## REFERENCES

- Bean C. R., Layard, P. R. G. and Nickell, S. J., "The Rise in Unemployment: a Multi-country Study," *Economica*, May 1986, suppl., 53, S1-S22.
- Bruno, M., "Aggregate Supply and Demand Factors in OECD Unemployment: An Update," *Economica*, May 1986, suppl., 53, S35-S52.
- and Sachs, J., *Economics of Worldwide Stagflation*, Cambridge: Harvard University Press, 1985.
- Goldthorpe, J. E., *Order and Conflict in Contemporary Capitalism*, Oxford: Oxford University Press, 1984.
- Rowthorn R. E., "Conflict, Inflation and Money," *Cambridge Journal of Economics*, September 1977, 1, 215-39.
- and Glyn, Andrew, "The Diversity of Unemployment Experience Since 1973," in S. Marglin, ed., *The Golden Age of Capitalism: Lessons for the 1990s*, forthcoming.
- OECD, *Economic Survey of Sweden*, Paris, 1984, 1985.
- , *Labour Force Statistics, Historical Statistics*, Paris, 1987.

# TAX POLICY AND INVESTMENT: A RECONSIDERATION<sup>†</sup>

## Investment, Financing Decisions, and Tax Policy

By STEVEN FAZZARI, R. GLENN HUBBARD, AND BRUCE PETERSEN\*

Studies of tax policy and corporate investment have been prominent in public finance and macroeconomic research. By integrating corporate income tax rates, investment tax credits, and the value of depreciation allowances into the "cost of capital," economists have analyzed the effects of taxes on capital spending. Most studies assume that firms respond to prices set in centralized securities markets, such as market interest rates of Tobin's  $q$ , and firms undertake all profitable investment projects. Firms choose the mix of finance among internal funds, debt, and new equity independently; the availability of finance does not limit investment. The implications for tax policy are clear: the marginal tax rate on returns from a new project matters for investment, not the firm's average tax burden on returns from its investments in place.

For firms that face imperfect markets for external finance, however, it is no longer sufficient to focus only on the cost of funds determined in centralized securities markets. In particular, if the cost of internal finance differs substantially from external finance for some firms, their investment depends on available cash flow. For these firms, the amount of earnings devoted to taxes—and therefore the *average* tax rate on returns

from existing projects—matters for investment, possibly along with incentive effects of marginal tax rates. We build on recent research that analyzes asymmetric information between firms and suppliers of external capital to examine the link between finance and investment, and, correspondingly, the role played by average as opposed to marginal tax rates in the investment process.<sup>1</sup>

Our approach emphasizes firm *heterogeneity*; some firms can obtain low-cost external funds to respond completely to signals from centralized securities markets, while internal finance constrains the investment of others. In Sections I and II, we consider a  $q$  investment model for two types of firms: (i) firms that face essentially no cost disadvantage of external finance, and (ii) firms that must pay a significant premium to raise funds from external sources because of information asymmetries. In both cases, we evaluate the relative influence of average and marginal tax rates on investment. Section III summarizes empirical evidence on the link between internal finance and investment across heterogeneous groups of firms. Section IV concludes and discusses implications for tax policy.

### I. The Supply of Finance and Investment

With perfect capital markets and no taxes, constraints on finance would never prevent firms from investing in projects with positive net present value. Firms would offset fluctuations in cash flow by issuing new debt or equity as needed to sustain their optimal capital accumulation programs. Dividend

<sup>†</sup>*Discussants:* Allen Sinai, Shearson Lehman Brothers/American Express Inc; Charles R. Hulten, University of Maryland; Roger Gordon, University of Michigan.

\*Department of Economics, Washington University, St. Louis, MO 63130; Department of Economics and Center for Urban Affairs and Policy Research, Northwestern University and NBER; and the Federal Reserve Bank of Chicago and Department of Economics, Northwestern University, Evanston, IL 60201; respectively. Hubbard acknowledges support from a John M. Olin Fellowship at the NBER.

<sup>1</sup>See our 1987 paper for a more formal exposition of these ideas and an extensive list of references.

policy is independent of investment. In equilibrium, firms' marginal  $q$  values are driven to unity, absent tax considerations.

If the personal tax system favors capital gains (taxed at an accrual-equivalent rate  $c$ ) over dividends (taxed at a rate  $\theta > c$ ), however, the after-tax cost of internal funds will be lower than that of external finance. If shareholders have no explicit preference for dividends, firms would exhaust internal finance before seeking external funds. Shareholders are indifferent between a dollar of retentions reinvested in the firm and taxed at rate  $c$ , and a dollar of dividends taxed at rate  $\theta$  if the shadow value of an additional unit of capital (marginal  $q$ ) is just  $(1 - \theta)/(1 - c) < 1$ . It is optimal for the firm to issue new shares only if marginal  $q$  exceeds unity. Thus, the tax system creates a "financing hierarchy" in which the threshold  $q$  value for marginal investment depends on the availability of internal finance.<sup>2</sup> Quantitatively, however, the tax advantage of internal finance was probably never particularly large, and the recent tax reform sharply reduced the personal tax advantage of capital gains income.

Asymmetric information about firms' prospects between firms and potential investors, however, can create a substantial cost differential between internal and external funds. This point is best illustrated in the case of new equity finance. Stewart Myers and Nicholas Majluf (1984) and Bruce Greenwald, Joseph Stiglitz, and Andrew Weiss (1984) explain how asymmetric information causes suppliers of new equity to demand large premia, or can eliminate the possibility of new equity financing all together. Suppose managers have better information than potential new shareholders about the true value of the firm. The true value will be revealed eventually, but new shares must be issued before that date, or the investment opportunity is lost—a realistic assumption, especially for firms in in-

dustries experiencing rapid technological advancement. A modified version of the market for "lemons" argument shows that firms may turn down some investment projects with positive net present values rather than issue new shares. In our earlier paper we show that the break-even  $q$  value for a new investment project is  $1 + \Omega > 1$ , where  $\Omega$  is the "premium" necessary to compensate new investors for the losses they incur from inadvertently funding lemons.

Firms that do not issue shares to finance projects with marginal  $q$  above one cannot necessarily substitute debt, because the asymmetric information problem is relevant for new debt as well. In some cases, asymmetric information causes lenders to impose a maximum debt-to-capital ratio.<sup>3</sup> Stiglitz and Weiss (1981) show how asymmetric information causes "credit rationing" for some borrowers. Lenders will evaluate internal equity or cash flow when making loans under asymmetric information, so that new debt finance depends on the availability of internal funds (see, for example, Ben Bernanke and Mark Gertler, 1987; and Charles Calomiris and Hubbard, 1987). Therefore, firms facing binding constraints in the equity market will likely face constraints in debt markets for similar reasons.

Figure 1 shows a financing hierarchy based on the combined effects of taxation and asymmetric information. Firms exhaust internal finance first and issue new shares only if the marginal project has a  $q$  of at least  $1 + \Omega$ . A firm with internal finance of  $R$  and an investment demand schedule of projects ranked by Tobin's  $q$  of  $D_1$ , would finance investment internally and pay some dividends. With investment demand  $D_2$ , it would exhaust internal finance, but not issue new shares. Only if a firm's investment demand schedule intersects the upper segment of the supply of finance schedule, as with  $D_3$  for example, will it issue new shares despite the

<sup>2</sup>See Alan Auerbach (1979), David Bradford (1981), and Mervyn King (1977). If dividends are valued for agency reasons, shareholders would discount retained funds, lowering the cost advantage of internal finance.

<sup>3</sup>It may be more realistic to assume that firms pay a rising marginal cost for new debt as investment increases rather than being strictly rationed. This change, however, does not materially affect the argument presented here.

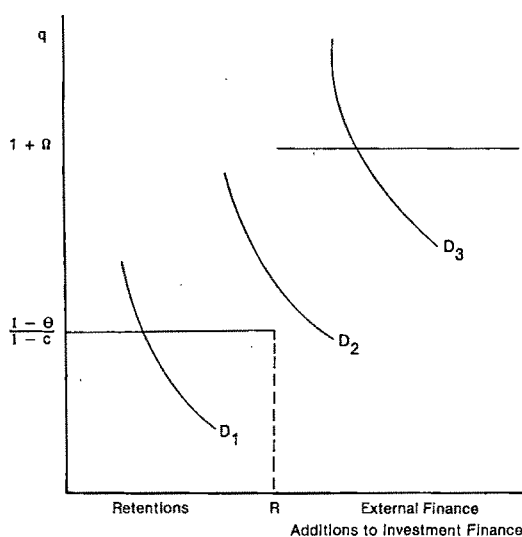


FIGURE 1. INVESTMENT AND FINANCING DECISIONS

cost disadvantages of external finance. When this cost disadvantage is significant, movements in internal finance can lead to fluctuations in investment spending.

Empirical evidence supports this view of financing behavior (see, for example, Myers, 1984). Philip Vijay Srinivasan (1986) shows that manufacturing corporations rely heavily on internal finance, particularly small corporations that are most likely to face asymmetric information problems. In addition, the average retention ratio of small corporations is very high and many of them pay no dividends at all for long periods of time. Many profitable small corporations exhaust internal finance, but do not borrow through long-term debt or issue new equity. This is consistent with a large lemons premium; firms that use all their internal finance may effectively be at a corner solution to their optimal investment finance problem.

## II. The Effect of Marginal and Average Tax Rates on Investment

Modern public finance focuses on how taxation affects the *marginal* cost of capital. In the  $q$  theory, changes in investment tax credits or depreciation allowances alter the equilibrium value of marginal  $q$  (see

Lawrence Summers, 1981, for example). In Figure 1, changes in tax incentives raise or lower the vertical intercepts of the internal and external finance segments. For firms with demand schedules like  $D_1$  or  $D_3$ , these changes affect investment.

This need not be true, however, for firms with investment demand schedules like  $D_2$  that operate at a corner solution.<sup>4</sup> Marginal changes in the cost of capital do not affect their investment. In this case, investment depends on the *average* tax on earnings from existing projects, not the marginal tax rate. Changes in the average tax burden expand or shrink the length of the internal finance (retentions) segment in Figure 1.

The discontinuity in Figure 1 at the point where internal finance is exhausted represents a limiting case for a firm with no access to marginal debt finance. If it can obtain new debt, the financing hierarchy would include an upward-sloping debt segment. The slope of this segment would indicate the expected marginal cost of financial distress. In this case, an increase in the firm's average tax burden would result in some substitution of debt for reduced internal finance, but at a higher cost.

We are not suggesting that marginal effective tax rates on new investment are not, in general, important. The point is, rather, that firms facing asymmetric information that exhaust their internal finance may have no low-cost substitute for internal funds at the margin. When the cost disadvantage of external finance is substantial, changes in average tax burdens have a pronounced effect on their investment behavior while marginal tax rates have relatively less impact. Again, firm heterogeneity is potentially important for understanding the channels through which tax policy influences investment. Furthermore, while our discussion has been cast in

<sup>4</sup>We abstract here from problems of tax losses and imperfect loss offsets. In our 1987 paper, we show that firms with investment most sensitive to internal finance are also most likely to experience losses, further weakening the effects of marginal incentives on investment.

terms of the  $q$  framework, similar logic applies to models emphasizing the "cost of capital." To the extent that empirical work relies on the cost of capital determined in centralized securities markets, investment spending in firms with large cost disadvantages of external finance will respond less to tax-induced changes in the cost of capital than investment in mature firms.

### III. Sensitivity of Investment to Internal Finance: Empirical Evidence

The  $q$  theory provides an empirical framework for analyzing the impact of changes in marginal investment tax incentives. To the extent that investment depends on internal finance and average tax rates, however, analyzing the effects of tax policy on  $q$  alone will not be sufficient. Recently, Andrew Abel and Olivier Blanchard (1986) found a significant role for profits in aggregate  $q$  investment equations, suggesting problems of aggregation or capital market imperfections. Our empirical approach addresses both of these issues.

To consider the effect of capital market imperfections for some firms arising from asymmetric information, we must rely on firm-level data. Our 1987 paper analyzes a sample of 421 manufacturing firms from 1969 to 1984. We identify differences in  $q$ , financing behavior, and investment across classes of firms defined by their *retention behavior*. If the cost disadvantage of external finance were small, retention behavior should contain little information about  $q$  or investment — firms would simply smooth fluctuations in cash flow with external finance. On the other hand, if information problems are important for growing firms, they will likely exhaust their internal funds and be on the margin of seeking outside finance, possibly at a corner solution. Their investment should vary substantially with fluctuations in cash flow.

Regression results reported in our earlier paper show that investment spending by growing firms paying zero or very low dividends over long periods displayed substantial "excess sensitivity" to movements in cash flow. Variations in cash flow alone for these firms explained a large proportion of the

variance in their investment-to-capital ratios.<sup>5</sup> Investment in more mature firms with high dividend payouts was not importantly influenced by cash flow.<sup>6</sup> This result illustrates the importance of firm heterogeneity both for the explanation of investment and the analysis of tax policy. In particular, changes in the average tax burden of firms that cannot obtain low-cost external finance may be of greater importance than incentive effects from changes in their marginal taxes.

The recent repeal of the investment tax credit (ITC) provides an illustration of this point. Standard models focus on the change this caused in the cost of capital or  $q$ . Repeal of the ITC is often considered especially significant for investment because of its subsidy effects. This may be true for mature firms. For firms at a corner solution on the financing hierarchy, however, the impact on internal finance of the repeal of the ITC may be more important than its effect on the cost of capital.<sup>7</sup> Therefore, in new, fast-growing industries where information problems will likely be most severe (high-technology industries, for example), the repeal of the ITC will have little special significance for investment relative to other aspects of tax reform that also affect these firms' average tax burden.

### IV. Implications for Tax Policy

The importance of internal finance for some firms' investment indicates potential pitfalls in policy analysis that considers only the impact of marginal tax rates on capital

<sup>5</sup> We divided firms into four classes according to their dividend payout ratios over a long period. The standard deviations of both investment and cash flow (relative to the capital stock) were four times larger in the lowest payout group than in the highest payout group.

<sup>6</sup> This result was especially striking when the standard  $q$  model was augmented to include not only cash flow, but also sales or "accelerator" effects.

<sup>7</sup> With the modified  $q$  model estimated in our 1987 paper and data on the cash value of the ITC across firms in 1985, we estimated that the cash flow effect of eliminating the ITC for firms that pay zero or low dividends was over *seven times* larger than the incentive effect through conventional channels. For mature firms, however, the incentive effect clearly dominates.

spending. As a *positive* matter, marginal investment incentives may have a weaker effect on investment than conventional models suggest, while average tax rates on profits from existing investments may be important. As a *normative* matter, however, it is difficult to design efficiency-enhancing policy reforms in this environment.

Tax policy can increase the *level* of investment within the framework that we outlined. For example, elimination of the corporate tax would stimulate investment in firms facing financing constraints, but would also confer a windfall gain on holders of old capital, with only an indirect feedback on investment for firms operating in perfect capital markets. An alternative "cash flow tax," much like a consumption tax, could be implemented by expensing investment and taxing the excess of profits over investment. This reform would lower average tax rates for growing firms investing all of their internal finance. For mature firms with internal cash flow in excess of investment, marginal incentives are preserved, though the average tax rate is high.

At least two factors mitigate the ability of such policies to increase *efficiency*. First, to the extent policymakers can distinguish project types no better than private financiers, the lemons problem remains. A second concern relates to agency issues (see Michael Jensen, 1986, for an overview). Policies that increase internal finance might encourage managers concerned, say, with corporate size as well as the value of shareholders' claims to overinvest. In equilibrium, however, overinvestment in low-marginal- $q$  projects should precipitate a takeover, with a profitable elimination of wasteful investment.

Our emphasis on information problems in markets for external finance may also shed light on the question of how tax policy affects dividend decisions. Intuitively, firms may pay dividends to overcome agency problems by limiting managers' discretion in making investments. Firms pay dividends until the marginal benefit matches the tax cost. But this framework does not explain the substantial cross-sectional variation in payout rates (see our earlier paper). For some firms that face external finance constraints, dividend

and investment decisions will not be independent. In this case, marginal dividends have an additional shadow cost representing investment displaced. Firms in this situation pay out less, not because their agency problems are less severe, but because their marginal cost of dividends from investment foregone can be substantial.

In summary, the importance of information-related imperfections in equity and debt markets suggests the need to reexamine the effect of internal finance and balance sheet positions on investment, and the corresponding implications for tax policy. In spite of the almost exclusive emphasis on marginal incentive effects of taxes, average tax burdens may be important for investment decisions in some firms, especially rapidly growing firms in industries with new technologies. Also, asymmetric information makes it unlikely that households can "pierce the corporate veil." Redistribution of tax burdens from corporations to households will influence the allocation of investment funds and the level of investment to the extent that firms face information-related financing constraints.

## REFERENCES

- Abel, Andrew B. and Blanchard, Olivier J., "The Present Value of Profits and Cyclical Movements in Investment," *Econometrica*, March 1986, 54, 249-73.
- Auerbach, Alan J., "Wealth Maximization and the Cost of Capital," *Quarterly Journal of Economics*, August 1979, 93, 433-46.
- Bernanke, Ben and Gertler, Mark, "Financial Fragility and Economic Performance," NBER Working Paper No. 2318, July 1987.
- Bradford, David, "Tax Incidence and Allocation Effects of a Tax on Corporate Distributions," *Journal of Public Economics*, April 1981, 15, 1-22.
- Calomiris, Charles W. and Hubbard, R. Glenn, "Firm Heterogeneity, Internal Finance and Credit Rationing," NBER Working Paper No. 2497, December 1987.
- Fazzari, Steven, Hubbard, R. Glenn and Petersen, Bruce C., "Financing Constraints and Corporate Investment," NBER Working Paper

- No. 2387, September 1987.
- Greenwald, Bruce, Stiglitz, Joseph E. and Weiss, Andrew**, "Information Imperfections in the Capital Market and Macroeconomic Fluctuations," *American Economic Review Proceedings*, May 1984, 74, 194-99.
- Jensen, Michael D.**, "Agency Costs of Free Cash Flow, Corporate Finance, and Takeovers," *American Economic Review Proceedings*, May 1986, 76, 323-29.
- King, Mervyn A.**, *Public Policy and the Corporation*, London: Chapman and Hall, 1977.
- Myers, Stewart C.**, "The Capital Structure Puzzle," *Journal of Finance*, May 1984, 39, 575-92.
- \_\_\_\_\_ and **Majluf, Nicholas S.**, "Corporate Financing Decisions When Firms Have Investment Information That Investors Do Not," *Journal of Financial Economics*, June 1984, 13, 187-220.
- Stiglitz, Joseph E. and Weiss, Andrew**, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, 71, 393-410.
- Srini Vasan, Philip Vijay**, "Credit Rationing and Corporate Investment," unpublished doctoral dissertation, Harvard University, October 1986.
- Summers, Lawrence H.**, "Taxation and Corporate Investment: A  $q$ -Theory Approach," *Brookings Papers on Economic Activity*, 1:1981, 67-127.

# Business Tax Policy, The Lucas Critique, and Lessons from the 1980's

By ROBERT S. CHIRINKO\*

In 1976, Robert Lucas published a devastating critique of the then current practice for quantifying the effects of alternative policies. He argued that, in formulating plans, economic agents necessarily look into the future, and thus the decision rules guiding their actions depend on parameters describing the expectations of future variables, as well as parameters of taste and technology. Lucas viewed economic policy as the selection of rules that generate the values of policy variables, rather than the selection of arbitrary sequences of policy variables. Thus, "any change in policy will systematically alter the structure of econometric models" (1981, p. 126), and the estimated coefficients in (the then current) consumption, wage-price, or investment models could not be considered structural, that is, invariant to alternative policy regimes. The important and damning implication for policy analysis is that these econometric relationships will prove unstable in precisely those situations in which they are called upon to analyze proposed policies.

This paper seeks to exploit the substantial variation in business tax policy during the 1980's to shed some light on the quantitative importance of the "Lucas Critique" (LC) and to evaluate the ability of explicit theory to enhance econometric analysis. In August of 1981, the economic agenda of the "Reagan Revolution" began to be put in place with the passage of the Economic Recovery Tax Act (ERTA). This was followed by the Tax Equity and Fiscal Responsibility Act of 1982,

the Deficit Reduction Act of 1984, and the Tax Reform Act of 1986. Other changes—such as the increase in defense spending and growth of and restrictions on the federal deficit—indicate that expectations of the current and future course of fiscal policy were altered radically during the Reagan Administration.

Maintained in this paper is the assumption that these changes in tax and other aspects of fiscal policy can be viewed as an unanticipated change in policy regime. In regard to the 1981 legislation, the *Economic Report of the President* claimed that "[T]his tax policy is a sharp break from the policies of the recent past. It reflects a different understanding of the way tax policy affects the U.S. economy" (1982, p. 109). Under the assumption that the 1980's witnessed a fundamental change in the stochastic process characterizing fiscal policy, econometric equations susceptible to the LC should become temporally unstable. Those models that successfully address the LC by isolating expectation parameters should remain stable.

The quantitative importance of the LC is evaluated by examining the stability of four different econometric models of business investment that to varying degrees are vulnerable to the LC: the "Neoclassical" model studied extensively by Robert Hall and Dale Jorgenson (1967), Robert Eisner and M. Ishaq Nadiri (1968), and their numerous collaborators, the  $Q$  model associated with James Tobin, the Vector Autoregressive model (VAR-S) advanced by Christopher Sims (1982), and a Hybrid VAR model (VAR-H) used to study investment by Robert Gordon and John Veitch (1986). Since the four models differ in the extent to which they are guided by explicit theoretical frameworks, these results will also be useful in assessing the impact of economic theory on econometric performance.

\*University of Chicago, Chicago, IL 60637. I thank Charles Hulten, Robert Lucas, and Allen Sinai for critical comments on a preliminary draft. All errors, omissions, and conclusions remain my sole responsibility.



### I. Four Econometric Models

The most frequently used econometric investment equation, and the one that was critiqued by Lucas, is the "Neoclassical" model. In this specification, the flow of investment is related to distributed lag (percentage) changes in output and the user cost of capital, where this latter term depends on the relative purchase price of new capital and the rates of taxation, interest, and depreciation.<sup>1</sup> Separate equations are estimated for equipment and structures, and are corrected for first-degree serial correlation in the residuals. As with all models studied in this paper, the dependent variables are scaled by their own capital stock and, to avoid issues of simultaneity, no contemporaneous variables are included. The distributed lag for user cost is longer than that for output, and they are constrained to lie along a third-degree polynomial with no endpoint restrictions.<sup>2</sup> These distributed lag coefficients can be interpreted as an amalgam of expectation and technology parameters. Insofar as the change in tax policy regime affects forecasts of future tax policy, the user cost coefficients will vary. In addition, if the regime change also affects the serial correlation properties of output, the other set of distributed lag coefficients will change as well.

In the wage of the LC, applied research has taken one of two directions. One research program uses explicit optimizing frameworks to develop models that are truly structural in the sense that their coefficients depend solely on parameters of taste and technology, which are invariant to changes in policy regimes. As discussed in my paper (1988), these models all follow from the same

analytic framework, differing only in the nature of the dynamic technology analyzed (for example, delivery lags, internal adjustment costs) and the manner in which unobservable expectations are related to observable variables. The  $Q$  model relies primarily on financial asset prices to solve this latter problem, and relates the flow of investment to a 10-quarter distributed lag of  $Q$ . (An alternative solution to the problem of unobserved expectations uses forecasting equations; see my 1988 paper.) Separate equations are estimated for equipment and structures, and are corrected for first-degree serial correlation in the residuals. Under this theory, the  $Q$  variable captures all the relevant information about expectations, and the estimated coefficients are related solely to the adjustment cost technology. Thus, a policy regime change will affect the observed  $Q$  variable, but the estimated coefficients are expected to remain unaltered.

An entirely different approach has been taken by Sims, who argues against the asymmetry in Lucas' model between the decision rules of private agents derived from an explicit optimization problem and the arbitrarily specified government policy rule. Under this view of policy, the changes in the 1980's do not reflect a new fiscal policy regime, but rather are interpreted as realizations from the same stable probability distribution characterizing purposeful behavior by both private and public agents. Sims models behavior through vector autoregressions (VAR), which impose few restrictions on the data and, under his view of policy, are expected to remain stable during the fiscal interventions of the 1980's. In the present study of investment, the VAR-S contains six variables: the level of output, the financial value of the firm, and a lagged dependent variable (all scaled by the capital stocks of equipment plus structures), the nominal interest rate, tax variables, and the relative price of investment goods (which constitute the user cost of capital but are entered separately and in levels). Each variable is lagged for twelve quarters, and the distributed lag coefficients are constrained to lie along a fourth-degree polynomial with no endpoint restrictions.

<sup>1</sup>A much more extensive discussion of the derivations underlying the estimated equations can be found in my papers (1986, 1988). The 1988 paper argues that all investment models can be derived from a general Neoclassical framework, hence the use of quotation marks around the label for Jorgenson's particular version of this model.

<sup>2</sup>The form and length of the lags were chosen after some *ad hoc* experimentation that sought to obtain significant sums of the coefficients. Note that the tests to be discussed in Section II were conducted after this preliminary experimentation had been completed.

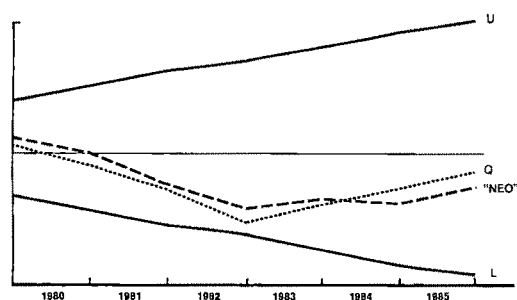
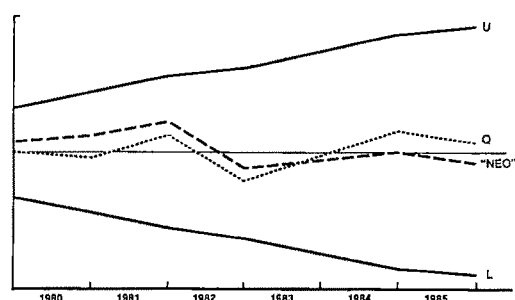
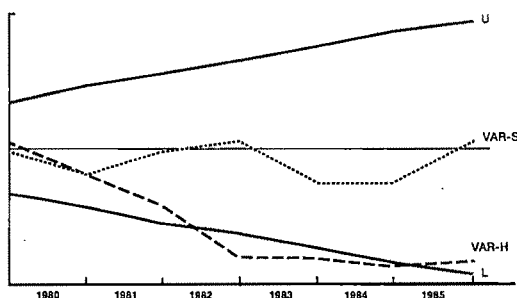
FIGURE 1. CUSUM STATISTIC: EQUIPMENT, "Neo,"  $Q$ FIGURE 3. CUSUM STATISTIC: STRUCTURES, "Neo,"  $Q$ 

FIGURE 2. CUSUM STATISTIC: EQUIPMENT, VAR-h, VAR-s

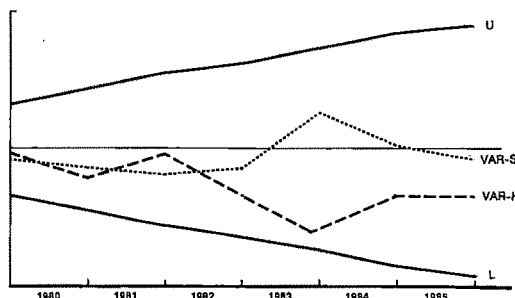


FIGURE 4. CUSUM STATISTIC: STRUCTURES, VAR-h, VAR-s

The fourth model combines the VAR approach with variable definitions from the Neoclassical and  $Q$  theories. In this Hybrid VAR framework (VAR-H) introduced by Gordon and Veitch, a model is estimated similar to the one above but with only the following four variables: the *levels* of output and the user cost of capital, a lagged dependent variable, and Tobin's  $Q$ . While a policy regime change in the 1980's will also destabilize the VAR-H model, it will be interesting to observe whether the additional information exploited by this approach has any bearing on the empirical results.

## II. An Assessment of the Lucas Critique

In order to examine equation stability during the 1980's, hence the quantitative importance of the LC, the cumulative sum of the recursive residuals (CUSUM) from each of the eight models is examined.<sup>3</sup> The recursive

residuals are calculated by estimating a model with quarterly data for the nonfinancial corporate sector for the period 1959:1 to 1979:4, then using this estimated coefficient vector and the explanatory variables for 1980:1 to calculate the one-step-ahead forecast error. With the sample for the regression extended by one period, this procedure is repeated sequentially for the 24 periods from 1980:1 to 1985:4. When the specification is adequate, the recursive residuals are distributed randomly, and CUSUM should not exhibit any persistent deviations from zero. The CUSUM statistic possesses the advantages that it has a known distribution and that no stand needs to be taken as to when the instability may have occurred, a particularly important feature because of anticipation effects and a divergence between expected passage, actual passage, and the effective date for tax legislation.

<sup>3</sup> The tests were introduced by R. L. Brown, J. Durbin, and J. M. Evans (1975), and are described succinctly in Andrew Harvey (1981, pp. 54-57; 148-54). After a first

draft of this study was completed, Lucas directed me to the doctoral dissertation of Jean-Marie Dufour (1979), who utilized recursive residuals to test the stability of a Hall-Jorgenson-Gordon investment model in the face of fluctuations in the investment credit during the 1960's.

The CUSUM statistics for the eight models are presented in Figures 1–4 with the upper (*U*) and lower (*L*) bands defining a confidence interval calculated at the 10 percent level. Except for the VAR-S models, there exists a noticeable downward trend beginning in 1981:3—the quarter in which ERTA became law—of five to six quarters for equipment (Figures 1 and 2) and seven to eight quarters for structures (Figures 3 and 4). A persistent upward trend in the VAR-S specification begins for both equipment and structures in the latter part of 1983. When this serial correlation is evaluated by a modified von Neumann ratio (Harvey, p. 156), it is found to be statistically significant for seven models (the exception being “Neoclassical” equipment). In conjunction with my assumption of a change in policy regime, these results highlight the parameter instability associated with the LC.

However, there is little evidence in favor of the quantitative importance of the LC. For all but the VAR-H equipment model, the CUSUM residuals are within the 10 percent confidence intervals.<sup>4</sup> Furthermore, as a percentage of the dependent variable, the one-step-ahead forecasts exceed 10 percent only twice (for VAR-S). For the “Neoclassical” and *Q* models, the error for the equipment or structures equations (considered separately) exceeds 6 percent only twice. Despite the turbulent fiscal climate of the 1980’s, these investment equations considered as a whole do not show any signs of important structural instability.

A striking pattern in the figures is that the CUSUM’s move together across models. A principal components analysis reveals that the first principal component accounts for over 75 percent of the variance of the recursive residuals for either the four equipment or four structures equations considered as a separate system. Since the investment models are based on different solutions to the

problem of unobservable expectations, this suggests that economywide shocks weigh more heavily in destabilizing investment than variations in expectation parameters.

The four models rely to varying degrees on explicit theory, with the *Q* model being the most rigorous, the “Neoclassical” model being theoretically correct in an environment of static expectations, the VAR-H exploiting some of the information contained in these two models, and the VAR-S being the least structured. (It should be noted that Sims has argued that the VAR model is fully consistent with restrictions imposed by dynamic economic theory; see Thomas Sargent, 1984, p. 408.) The figures are not useful in assessing forecast accuracy because they are sensitive to the sign of the errors; a string of small negative errors will be more noticeable than errors that are much larger but of opposite sign. The mean squared error is the preferred measure of forecast accuracy, and the values for (equipment/structures) are as follows: *Q* (29/8), “Neoclassical” (26/7), VAR-H (24/13), and VAR-S (65/17). While no model emerges as superior, the relatively unstructured VAR-S model clearly performs poorest.

### III. Conclusion

This paper has assessed the stability of equipment and structures investment equations in the face of the volatile fiscal environment of the 1980’s. The CUSUM test was able to identify the instability following from the Lucas Critique, but its quantitative impact was not of major importance. This conclusion was especially evident in the Neoclassical and *Q* models, which are based on optimizing frameworks. While the model-building strictures following from the Lucas Critique do not appear to be of critical quantitative importance, the results presented in this study suggest that empirical work is more likely to be successful if guided explicitly by economic theory.

### REFERENCES

- Brown, R. L., Durbin, J. and Evans, J. M.,  
“Techniques for Testing the Constancy of

<sup>4</sup>Note that the statistical significance of the CUSUM test is compromised in an unknown manner by the estimated value of the serial correlation parameter and the presence of lagged dependent variables. However, Brown et al. regarded these bounds as “yardsticks against which to assess the observed sample path rather than providing formal tests of significance” (p. 155).

- Regression Relationships Over Time," *Journal of the Royal Statistical Society, Series B*, 1975, 37, 149-92.
- Chirinko, Robert S., "Business Investment and Tax Policy," *National Tax Journal*, June 1986, 39, 137-55.
- \_\_\_\_\_, "Will 'The' Neoclassical Theory of Investment Please Rise?: The General Structure of Investment Models and Their Implications for Tax Policy," in J. M. Mintz and D. D. Purvis, eds., *The Impact of Taxation on Business Investment*, forthcoming, 1988.
- Dufour, Jean-Marie, "Methods for Specification Error Analysis with Macroeconomic Applications," unpublished doctoral dissertation, University of Chicago, 1979.
- Eisner, Robert and Nadiri, M. Ishaq, "Investment Behavior and Neoclassical Theory," *Review of Economics and Statistics*, August 1968, 50, 369-82.
- Gordon, Robert J. and Veitch, John M., "Fixed Investment in the American Business Cycle, 1919-83," in R. J. Gordon, ed., *The American Business Cycle: Continuity and Change*, Chicago, London: University of Chicago Press, 267-335.
- Hall, Robert E. and Jorgenson, Dale W., "Tax Policy and Investment Behavior," *American Economic Review*, June 1967, 57, 391-414.
- Harvey, Andrew C., *The Econometric Analysis of Time Series*, New York: Wiley & Sons, 1981.
- Lucas, Robert E., "Econometric Policy Evaluation: A Critique," in K. Brunner and A. Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19-46; reprinted in *Studies in Business-Cycle Theory*, Cambridge: MIT Press, 1981, 104-30.
- Sargent, Thomas, J., "Autoregressions, Expectations, and Advice," *American Economic Review Proceedings*, May 1984, 74, 408-15.
- Sims, Christopher A., "Policy Analysis with Econometric Models," *Brookings Papers on Economic Activity*, 1:1982, 107-64.

# Investment Tax Incentives and Frequent Tax Reforms

By ALAN J. AUERBACH AND JAMES R. HINES, JR.\*

In the uncertain business of planning for U.S. corporate investment, one of the few reliable forecasts one can make is that the tax law will change before any new investment outlives its usefulness. While the Tax Reform Act of 1986 mandates an unusually dramatic reform in the structure of business taxation and the incentive to invest, the simple fact that Congress chose to alter in 1986 the tax treatment of new investments is hardly surprising. Earlier in the 1980's, Congress changed investment incentives with new tax legislation in 1981, 1982, 1984, and 1985, and over the period 1953-85 made such changes in 16 different years.

The willingness, indeed, eagerness, of the U.S. government to amend the rate at which it taxes new investments seems likely to have substantial consequences for investor incentives. By far the bulk of an investor's return comes in years subsequent to the year in which new plant and equipment is put in place. Tax reforms affect investor returns not only by changing the amount of money owed the government, but also by encouraging or discouraging competing future investment and thereby changing levels of before-tax future earnings. For example, the knowledge that Congress plans to introduce a large investment tax credit in two years seems likely to depress investment this year and next, since the investment wave two years hence will be expected to drive down the return to any capital already in place when it starts.

Despite the frequency of tax changes and their potential importance to investors, al-

most all of the analysis of tax-based investment incentives follows the seminal work of Dale Jorgenson (1963) in assuming investors never anticipate any tax changes. Recent examples include Auerbach (1983) and Mervyn King and Don Fullerton (1984). The U.S. Treasury Department in its tax reform proposal (1984) analyzed investment incentives in each year of its phased-in reform package under the assumption that investors never anticipate the sequence of tax changes that is explicitly part of the reform. In this paper, we depart from this approach by analyzing the historical pattern of U.S. corporate investment incentives over the period 1953-86, incorporating the feature of investor awareness that next year's tax code may not be the same as this year's.

## I. Anticipated Tax Reforms

In order to analyze the impact of expected future tax changes, it is necessary to understand over which tax variables investors form expectations. The tax law as written contains thousands of provisions affecting corporate investment. Even restricting attention just to the statutory tax rate, the investment tax credit, and the present value of depreciation allowances, leaves a problem of great complexity, since congressional choice of the level of one variable is surely conditioned by the chosen levels of the others.

There is, furthermore, a relevant issue of the extent to which the government's "choice" of a particular tax variable is truly volitional. In roughly half of the postwar years, the government did not make substantial legislative changes in any of the three variables listed. Yet even without specific congressional action, the *ex ante* value of depreciation allowances available on new investments varies from year to year with movements in expected inflation and real interest rates. And while these movements automatically affect the level of investment

\*University of Pennsylvania, Philadelphia, PA 19104 and NBER, and Woodrow Wilson School, Princeton University, Princeton, NJ 08544 and NBER, respectively. We thank Kevin Hassett for research assistance, Stephen Goldfeld and Roger Gordon for helpful comments, and the NSF, grant SES86-17495 and University of Pennsylvania Institute for Law and Economics for financial support.

tax incentives, they also are likely to be correlated with changes in general economic conditions, such as unemployment or GNP growth. Hence it may be the case that during periods of no tax changes, Congress was actually permitting automatic features of the tax system to set investment incentives at acceptable levels. Or, it could be that in the years in which no new tax laws appeared, Congress would have chosen to set tax rates at different levels, but was for some reason prevented from making any changes that year.

In order to model anticipations of future tax reforms, we assume investors to expect tax changes based on a model of government choice in which the government reveals its desired tax levels only in those years that it enacts new tax laws. Thus, the 16 years in which the corporate tax law changed over the period 1953–85 afford us 16 glimpses of the outcome of the government's desired tax function. We take the probability that a new tax law will be enacted to be exogenous. (This is a very simple and perhaps inadequate specification; we are currently pursuing work based on alternative assumptions.) If no new law appears, then, of course, the preexisting tax law applies to new investments, with possibly new incentives due to changes in inflation and interest rates.

There still remains the issue of specifying the particular future tax variable investors predict when making their decisions. Following Jorgenson and subsequent authors, we assume that a firm expecting no future tax changes will set its marginal product of capital equal to

$$q(\rho + \delta) \frac{(1 - k - \tau z)}{(1 - \tau)}$$

where  $q$  is the relative price of capital goods,  $\rho$  is the real discount rate,  $\delta$  is the geometric rate of economic depreciation,  $k$  is the investment tax credit,  $\tau$  is the corporate tax rate and  $z$  is the present value of depreciation allowances per dollar invested. The cost of capital, then, is directly affected by the ratio  $(1 - k - \tau z)/(1 - \tau)$ , and it is this ratio which we assume Congress to peg in making its tax choices.

This choice of tax specification raises an issue that bears on the appropriateness of different sample periods for estimating the government's choice function. This model embodies the assumption that changes in the statutory tax rate,  $\tau$ , apply only to new investments. Of course this is not the case, but in defense of our procedure, the statutory corporate tax rate changed infrequently over the period 1953–1985, and moved very little when it moved at all: it ranged from a high of 52.8 percent to a low of 46 percent. The Tax Reform Act of 1986 departed widely from this pattern by introducing a phased reduction of the statutory rate to 34 percent, and for that reason we leave 1986 out of our estimating sample.

Denoting the ratio  $(1 - k - \tau z)/(1 - \tau)$  for aggregate corporate investment in year  $t$  by  $Tax_t$ , (with  $z$  calculated using an assumed real interest rate of 4 percent and static inflation expectations), we estimated several equations with explanatory variables suggested by our view of the factors likely to affect policy choices. The best fit was obtained with

$$\begin{aligned} (1) \quad Tax_t = & 1.90 - 0.045 UNEM_t \\ & (0.16) \quad (.006) \\ & - 0.021 YGRO_t - 0.017 REALR_t \\ & \quad (.004) \quad (.004) \\ & - 1.94 IGNP_{t-1} \\ & \quad (.89) \\ n = 16, \bar{R}^2 = .88, SEE = 0.036 \end{aligned}$$

where  $UNEM$  is the unemployment rate,  $YGRO$  is the real growth rate of GNP from the previous year,  $REALR$  is the real interest rate on short-term commercial paper (using the GNP deflator to measure inflation), and  $IGNP$  is the ratio of real investment to GNP. Standard errors are in parentheses.

Even allowing for the small number of degrees of freedom, all the variables except the last are significant at a 95 percent confidence level, and  $IGNP_{t-1}$  fails the 95 percent test only just barely. Equation (1) is characterized by a surprisingly good fit for a time-series regression with a nontrending de-

pendent variable, and in fact (1) predicts, out of sample, a tax reform in 1986 very similar to the change Congress enacted (see below). Neither current nor lagged inflation entered significantly in (1), indicating that Congress set tax rules with a view toward undoing the effects of changing inflation rates. Coefficients on the unemployment and real interest rates indicate that investment incentives are set in a countercyclical manner. However, the remaining coefficients have the "wrong" sign for that interpretation—and the whole notion of "countercyclical" tax policy using investment incentives has been questioned by Robert Lucas (1976), and by Finn Kydland and Edward Prescott (1977), who specifically discussed the impact on private behavior of government attempts to use the investment tax credit as a stabilization tool.

## II. Model

In order to analyze the impact of current tax changes and anticipated future tax reforms on investment incentives, it is necessary to construct a model of dynamic firm behavior. In particular, one must pay close attention to the specification of the costs firms face as they vary their investment levels. It seems quite reasonable empirically to employ a model in which rapid adjustment is costly, but the introduction of adjustment costs introduces a number of complications.

For our purpose in this paper, we employ a discrete-time variant of the model analyzed in Auerbach (1986). We assume the firm to maximize the expected present value of its after-tax cash flows:

$$(2) \quad V_t = E_t \left[ \sum_{s=t}^{\infty} (1+r)^{-(s-t)} \times \left\{ \frac{(1-\tau_s)p_{s+1}F(K_s)}{1+r} - p_s \left( 1 + \frac{1}{2}\phi I_s \right) I_s (1-k_s - \Gamma_s) \right\} + A_t \right]$$

where  $E_t$  is the expectations operator at time  $t$ ,  $r$  is the nominal discount rate,  $\tau_s$  is the statutory tax rate at time  $s$ ,  $p_s$  is the price

level for output at time  $s$ , and  $F(\cdot)$  is the firm's production function which exhibits decreasing returns with respect to capital,  $K$ . (One can also interpret  $F(\cdot)$  as a reduced form of a constant returns production function with levels of other factors chosen optimally.) The parameter  $\phi$  reflects investment adjustment costs, which are assumed to be capitalized and depreciated for tax purposes.  $k_s$  is the investment tax credit available in period  $s$ , while  $\Gamma_s$  is the present value of future depreciation allowances times future statutory tax rates (hence  $\Gamma = \tau z$  if  $\tau$  is expected to be constant).  $A_t$  is predetermined (though perhaps uncertain) in year  $t$ ; it is the value of financial attributes (such as future depreciation allowances on old investments) the firm cannot affect.

Maximization of (2) over the choice of investment (and hence the capital stock) in each year yields a first-order condition that (assuming that  $r$ ,  $\delta$ , and the inflation rate  $\Delta p/p$  are small), can be approximated by

$$(3) \quad \frac{F'(K_t)(1-\tau_t)}{1-k_t-\Gamma_t} = q_t(\rho + \delta) + \frac{E_t \Delta [q_t(1-k_t-\Gamma_t)]}{1-k_t-\Gamma_t}$$

in which  $\rho = r - \Delta p/p$  is the real discount rate,  $q_t$  is the pre-tax marginal cost of an additional piece of capital (inclusive of adjustment costs) relative to the price of output, and the operator  $\Delta$  denotes changes from one year to the next. If the second term on the right side of (3) were zero, so that investors expect no changes in the after-tax relative price of new capital between this year and next, then the formula implies that the after-tax cost of capital is exactly the same as that cost which emerges in the standard Jorgenson-type framework.

But, in general, investors will not expect the change in  $[q_t(1-k_t-\Gamma_t)]$  to be zero from year to year. The after-tax marginal cost of capital will be expected to change either through tax changes, or through changes in investment levels which affect marginal adjustment costs. And naturally the level of investment is itself a function of current and expected future tax policy. Hence

a consistent analysis of forward-looking investment tax incentives should incorporate not only the tax treatment of current investments, but also the effect of the tax law on the current level of marginal adjustment costs and expected changes in marginal adjustment costs.

Following Auerbach, it is possible to derive a fairly simple expression for the combined effect of taxes on marginal investment incentives. (Details are available from the authors on request.) Doing so requires some approximations, such as linearizing the model around steady-state values of investment and the capital stock, and it also requires a specification of the nature of anticipated tax changes. For the purposes of this solution, we assume investors to anticipate that the government will introduce (potentially) new values of  $(k + \Gamma)$  at some date, but are uncertain about the timing of their adoption. That is, investors at time  $t$  observe current tax parameters and also form expectations of the values the parameters would take in a tax reform package, and investors anticipate the probability of passage of a reform measure to be constant at a rate  $\pi$  per year. Investors expect the tax reform, once adopted, to be the true final resting spot for the tax system, with no further tax changes to follow.

Given this specification of the model, the cost of capital that determines the capital stock takes the form (suppressing subscripts):

$$(4) \quad q \left[ \frac{(\rho + \delta)(1 - k - \Gamma)}{1 - \tau} + \frac{\Delta(k + \Gamma)}{1 - \tau} \cdot \frac{[\lambda - (\rho + \delta)]\pi}{\lambda + \pi} \right]$$

where  $(1 + \lambda)$  is the unstable root of the second-order difference equation describing the evolution of the capital stock in response to tax changes.  $\lambda$  lies in the interval  $[(\rho + \delta), \infty]$  and approaches infinity as adjustment costs vanish. Use of (4) permits changes in the cost of capital to be decomposed neatly into two pieces: the first term, which is standard (see expression (1)) and the second term, which is the effect of future tax changes through  $\pi$  and the *average* value that tax variables are expected to take.

The cost of decomposing investor incentives in this way is that the model we use is somewhat stylized. Investors expect the statutory tax rate not to change, and anticipate any future changes in  $k$  and  $\Gamma$  to be permanent. This permanent change is adopted at a constant hazard rate  $\pi$ . While none of these assumptions is required in order to solve this model, their use greatly simplifies the problem at little likely cost of changing the results. The same evaluation appears to apply to the specification of the adjustment cost function in (2). Here a central issue is whether the cost of adjusting the capital stock is properly specified as a function of the rate of investment, or as a function of the rate of investment *relative to* the size of the capital stock. Since a growing economy exhibits secular growth in the former, it may be more reasonable to use the latter specification. Unfortunately, the phenomenon of ratio adjustment costs requires a different measure of economic depreciation from that commonly used (see Andrew Abel, 1983 and Auerbach, 1986). Auerbach (1986) and our paper (1987) analyze investment tax incentives with ratio adjustment costs. This complication would make our calculations difficult to compare with other studies that ignore adjustment costs altogether. Therefore, we present results which emerge from the model in (2) with level adjustment costs. (Calculations performed with ratio adjustment costs did not differ qualitatively from those reported in this paper, and are available from the authors.)

### III. Results

It is convenient to summarize investment tax incentives with effective tax rates, which measure marginal wedges between the gross and net of tax returns to capital. In the case that the tax system is expected to change, this rate may be interpreted as that constant tax rate on true economic income which would yield the current level of investment (see Auerbach, 1986).

We measure effective tax rates for aggregate nonresidential corporate investment using expression (4), with the expected tax change set to zero for the case of myopic expectations. The calculations also require



specification of several parameters. We set  $\delta = .0704$  (based on calculations in our paper, 1987), and  $\rho = .04$ . For "high-adjustment-cost" simulations, we choose a value of  $\phi$  that, when multiplied by the steady-state capital stock around which the approximation is taken yields a marginal adjustment cost equal to that imposed by a proportional adjustment cost model with a corresponding quadratic parameter of 20. This choice implies extremely slow adjustment, but is nonetheless somewhat lower than many empirical estimates (for example, see Lawrence Summers, 1981). For "low-adjustment-cost" simulations, we set  $\phi$  to correspond to a ratio adjustment cost parameter of .5. The root  $\lambda$  depends on  $\rho$ ,  $\delta$ ,  $\phi$ , and the local elasticity of the marginal product of capital with respect to the capital stock (details available on request). For this parameter, we use a value of .65, which is reasonable given its interpretation in the Cobb-Douglas production function as the labor share of *gross* output. This yields values for  $\lambda$  of .128 in the high-adjustment-cost case and .409 in the low-adjustment-cost case.

Finally, we must specify the probability of change perceived by investors and that tax system expected to be adopted if a change occurs. We consider two specifications of probability. The "variable reform probability" sets each year's probability to the fraction of the previous 5 years in which a tax change occurred, while the "constant reform probability" specification sets each year's probability of .5, the approximate fraction of sample years in which taxes changed. In all simulations, the potential tax reform expected in each year is that which equation (1) predicts given that year's economic variables.

Table 1 summarizes our findings. Column 1 presents effective tax rates as conventionally measured, that is, under the assumptions of no adjustment costs and no future changes in any relevant variables, including tax variables. Columns 2 and 3 present effective tax rates in the presence of large adjustment costs, and while these rates are not identical to standard measures of effective tax rates, the difference is quite small. This similarity arises because future investment tax incentives are of little importance to firms that

TABLE 1—EFFECTIVE TAX RATES ON CORPORATE INVESTMENT, 1953–86

Year	(1)	High Adjustment Costs		Low Adjustment Costs	
		(2)	(3)	(4)	(5)
1953	55.3	55.8	55.3	60.7	55.3
1954	49.5	49.2	49.5	45.4	49.5
1955	52.1	52.7	52.6	58.8	56.3
1956	54.4	55.1	54.9	61.2	58.7
1957	54.8	55.5	55.3	61.7	59.2
1958	50.9	51.5	51.3	56.7	54.5
1959	52.6	53.9	53.6	64.5	60.5
1960	50.6	51.7	50.6	60.6	50.6
1961	48.4	49.6	48.4	59.4	48.4
1962	37.7	37.6	37.7	36.0	37.7
1963	36.4	36.2	36.3	34.5	35.2
1964	34.3	34.2	34.2	32.5	33.3
1965	34.5	33.9	34.0	27.5	28.3
1966	37.7	37.5	37.5	35.4	35.2
1967	44.5	44.6	44.6	45.4	45.4
1968	46.0	45.9	45.9	44.8	44.7
1969	47.5	47.5	47.5	46.9	46.8
1970	50.8	51.2	51.3	55.5	56.3
1971	49.0	49.6	49.7	55.2	56.3
1972	33.7	33.2	33.1	27.4	25.3
1973	37.5	37.7	37.8	40.6	41.5
1974	42.6	42.5	42.5	41.0	40.6
1975	43.3	43.3	43.3	43.8	43.8
1976	33.2	33.0	33.0	30.4	30.2
1977	34.7	34.6	34.6	34.1	34.1
1978	37.8	38.6	38.5	46.0	43.0
1979	37.4	37.7	37.6	41.3	39.8
1980	38.2	38.7	38.7	44.1	43.6
1981	26.1	26.2	26.2	27.3	26.9
1982	17.6	17.0	17.0	9.5	10.4
1983	10.3	10.5	10.5	12.8	13.0
1984	12.7	13.6	13.6	21.9	22.6
1985	12.7	12.0	12.0	4.3	3.6
1986	9.2	6.7	6.5	-31.7	-45.0

Notes: Shown in percent. Col. 1 presents conventional myopic ETR and cols. 2 and 4 present constant reform probability; cols. 3 and 5 present variable reform probability for the respective high and low adjustment costs.

feel locked into today's investments by steep costs of varying their investment rates. Hence adjustment costs at the levels estimated by some authors make anticipated tax policy unimportant to contemporary investment decisions.

Columns 4 and 5 contain estimated effective tax rates in the presence of low but nonzero adjustment costs. These estimates exhibit considerably more variability than conventional effective tax rates, as investors typically anticipate actual movements in investment incentives and understand when changes come that the legislature often overshoots its mark. Examples of these anticipations include the periods 1971–73, 1980–82, and 1984–86. The scenario with variable reform probabilities shows this effect dramati-

cally during periods such as the early 1970's and mid-1980's, when rapid-fire tax reforms make investors very sensitive to the government's desired policy since it is likely to be enacted soon. By 1986, investors were so certain that a tax change was coming, and that the change would rob the tax system of many of its investment incentives, that contemporaneous investment incentives looked extremely attractive by comparison. (Although in the event the investment tax credit was repealed retroactive to January 1, 1986.)

Much more research is necessary in order to identify the full impact of anticipated tax reforms on investment incentives. Anticipations of future economic conditions and government policy responses can be modeled in a richer environment that incorporates nonstatic expectations, rational updating of investors' anticipation function, and the endogeneity of future economic variables to tax changes. Anticipated tax policy would be likely to assume more importance in a model that disaggregated investment by sector or by asset type, since these breakdowns would capture the very large historical differences and movements in the taxation of equipment and structures investments. But even without a more detailed model, it is clear that anticipated policy is important only if investment is relatively flexible.

## REFERENCES

- Abel, Andrew B., "Tax Neutrality in the Presence of Adjustment Costs," *Quarterly Journal of Economics*, November 1983, 98, 705-712.
- Auerbach, Alan J., "Corporate Taxation in the United States," *Brookings Papers on Economic Activity*, 2:1983, 451-505.
- \_\_\_\_\_, "Tax Reform and Adjustment Costs: The Impact on Investment and Market Value," NBER Working Paper No. 2103, 1986.
- \_\_\_\_\_, and Hines, James R., Jr., "Anticipated Tax Changes and the Timing of Investment," in Martin Feldstein, ed., *The Effects of Taxation on Capital Accumulation*, Chicago: University of Chicago Press, 1987.
- Jorgenson, Dale W., "Capital Theory and Investment Behavior," *American Economic Review Proceedings*, May 1963, 53, 247-59.
- King, Mervyn A. and Fullerton, Don, *The Taxation of Income from Capital*, Chicago: University of Chicago Press, 1984.
- Kydland, Finn E. and Prescott, Edward C., "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, 85, 473-93.
- Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique", in Karl Brunner and Allan H. Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19-46.
- Summers, Lawrence H., "Taxation and Corporate Investment: A  $q$  Theory Approach," *Brookings Papers on Economic Activity*, 1:1981, 67-127.
- U.S. Department of the Treasury, *Tax Reform for Fairness, Simplicity, and Economic Growth*, Washington, November 1984.

# TECHNOLOGICAL INNOVATION AND PRODUCTIVITY CHANGE IN JAPAN AND THE UNITED STATES<sup>†</sup>

## Productivity and Economic Growth in Japan and the United States

By DALE W. JORGENSON\*

During the period from 1960 to 1973, the economic growth rate in Japan was at the rate of 10 or 11 percent per year. Japan was not the only country that grew rapidly during that period. France and Germany grew at 5.9 and 5.4 percent per year between 1960 and 1973 and Italy grew at 4.8 percent per year. Even the United Kingdom grew at a respectable 3.8 percent per year. The United States grew at 4.3 percent per year during this period. To fill out the roster of the seven major industrialized countries, Canada grew at 5.1 percent per year.<sup>1</sup>

After the first oil crisis in 1973, and even more so after the second oil crisis in 1978–79, there was a dramatic decline of economic growth among industrialized countries. Growth in the OECD countries dipped to 2.6 percent per year between 1973 and 1979. Japanese growth dropped from the double-digit levels of the 1960's and the early 1970's to 3.8 percent per year from 1973 to 1979. In the United States, the growth rate dropped to slightly above the OECD average at 2.8 percent per year. The rate of economic growth in Germany dropped to 2.4 percent and in France to 3.1 percent. In every major

industrialized country there was a precipitous fall in the rate of economic growth.

The sources of economic growth in Japan and the United States over the whole period from 1960 to 1979 are given in Table 1. If we compare Japan and the United States during the period 1960–79, we see that the growth of output over the whole period was 8.3 percent in Japan and only 3.5 percent in the United States. We can allocate this growth in output in the two countries among its three sources, namely, the contribution of capital input, the contribution of labor input and the rate of technical change. By far, the most important contributor to economic growth in both countries is the growth of capital input. This growth source accounts for about 5 percentage points of the Japanese economic growth rate and about 1.5 percentage points of the U.S. economic growth rate. This amounts to 60 percent of Japanese growth and 40 percent of U.S. growth.

Labor input in the two countries is a major contributor to economic growth, accounting for 1.5 percent of the Japanese growth rate and 1.2 percent of the U.S. growth rate. The rate of technical change is an important contributor as well, at nearly 2 percent in Japan and 0.7 percent in the United States. I conclude that by far the most important contributor to economic growth in the two countries is the growth of capital input. The relative importance of capital input is much greater in Japan than in the United States.

Focusing attention on the period from 1973 to 1979 after the energy crisis, we can see that capital input retained its lead as a source of economic growth in both countries. However, the decline in the growth of

<sup>†</sup>*Discussants:* John W. Kendrick, George Washington University; J. Randolph Norsworthy, Rensselaer Polytechnic Institute; Rolf R. Piekarz, National Science Foundation.

\*Harvard University, Cambridge, MA 02138.

<sup>1</sup>Comparisons of patterns of economic growth in industrialized countries are given by Laurits Christensen, Diane Cummings, and myself (1980, 1981). Comparisons between Japan and the United States are given by myself and Mieko Nishimizu (1978), myself with Masahiro Kuroda and Nishimizu (1987), and myself with Sakuramoto, Yoshioka, and Kuroda (1988).

TABLE 1—SOURCES OF ECONOMIC GROWTH,  
JAPAN AND THE UNITED STATES, 1960–79

	1960–79		1973–79	
	Japan	U.S.	Japan	U.S.
<b>Average Annual Growth Rate</b>				
Net Output	0.083	0.035	-0.038	0.028
Capital Input	0.096	0.040	0.060	0.038
Labor Input	0.031	0.020	0.015	0.017
<b>Annual Rate of Contribution to Growth</b>				
Capital Input	0.050	0.015	0.029	0.014
Labor Input	0.015	0.012	0.008	0.011
Technical Change	0.020	0.007	0.001	0.003
Quality Change of Capital Input	0.018	0.004	0.005	0.003
Quantity Change of Capital Input	0.032	0.012	0.024	0.011
Quality Change of Labor Input	0.010	0.002	0.005	0.001
Hour Worked Change	0.005	0.010	0.004	0.010
Weighted Average of Sector Technical Change	0.007	0.004	-0.012	-0.007
<b>Contribution of Allocational Changes</b>				
Net Output	0.004	0.002	0.014	0.009
Capital Input	0.009	0.001	0.005	-0.000
Labor Input	0.001	0.000	-0.005	0.002

Sources: Myself with F. M. Gollup and B. M. Fraumeni (1988) and myself with M. Kuroda and M. Nishimizu (1987).

capital input in Japan was much greater than in the United States. The contribution of capital dropped from 5.0 percent in Japan to 2.8 percent in the period 1973–79. The contribution of labor declined as well, from 1.5 to 0.8 percent, and the rate of technical change dropped from approximately 2 percent to a mere one-tenth of 1 percent, 0.13 percent to be more precise, during the period from 1973 to 1979.

If we consider the corresponding figures for the United States, we see that there was an almost negligible decline in the contribution of capital input from 1.5 to 1.4 percent per year. The same is true of the contribution of labor. Therefore, the impact of the oil crisis on U.S. economic growth has to be traced to the decline in the rate of technical change, the so-called “unexplained residual,” which declined from 0.7 to 0.3 percent per year.

During short periods of time, it is possible for output growth to exceed the sustainable level by increasing the proportion of the

national product devoted to capital formation. That is the mechanism at work in the very rapid growth in the Japanese economy during the 1960's and the early 1970's. Referring again to Table 1, capital input grew 1.3 percent more rapidly than output in Japan during the period 1960–79. This was the consequence of the increase in the proportion of the national product that was devoted to capital formation.

Was the decline in the growth of output that took place in the period from 1973 to 1979 due to the fall in the growth of capital? Since capital is so important to Japanese economic growth, this is a potential explanation of the slowdown. In Table 1 we see that the decline in output was 4.5 percent. But capital declined only 3.6 percent. Therefore, after the energy crisis as well as before, the growth rate of capital input was higher than that of output. Rather than causing the slowdown, the growth of capital after the energy crisis contributed to the continued growth of output at unsustainable levels. My first conclusion is that the decline in the growth rate of capital is not the cause of the slowdown in Japanese economic growth.

Turning to the United States, we see that the output growth rate declined by about six-tenths of 1 percent while the capital growth rate declined by only two-tenths of 1 percent. Despite the fact that capital is the most important source of U.S. growth, the decline in the growth rate of capital was not the cause of the slowdown in the growth of output. In the United States as in Japan the growth of output was maintained at unsustainably high levels during the period after the energy crisis. The growth of capital did not account for the slowdown that took place in the United States.

There was a decline in the growth rate of labor input in Japan from an average of 3.1 percent for the period 1960–79 to 1.5 percent for the period 1973–79. If we consider the period 1973–79, we see that hours worked have continued to grow in Japan, but that the upgrading of the labor force has declined by about 50 percent. Labor quality change is a very important growth source and is part of the story of the slowdown in economic growth in Japan. I have now identified one factor that is clearly responsible for part of

the decline in the growth in Japan—the decline in the change of quality of labor input.

The contribution of labor quality in the United States dropped from 0.22 to 0.06 percent between the period 1960–79 and the period 1973–79. Hours worked in the United States have grown at rates almost double those in Japan throughout the period 1960–79. This remained the case during the period 1973–79. Hours worked grew even more rapidly during the period after the energy crisis than before. My second conclusion is that the reduction in the rate of upgrading of the labor force was an important factor in the decline of economic growth in both countries.

Finally, let us turn our attention to the rate of technical change. The decline from 1960–79 to the subperiod 1973–79 was 1.9 percent in Japan and 0.4 percent in the United States. It is clear the decline in the rate of technical change must play the predominant role in explaining the slowdown. The next question is: how is it possible to link the rate of technical change to energy prices? There is an element of truth in the idea that the growth of output at the aggregate level cannot be traced directly to the change in energy prices, since energy itself is a small proportion of aggregate output.<sup>2</sup> This is true in both Japan and the United States. However, this point of view ignores the fact that aggregate growth is the result of the growth of individual industrial sectors.

At this point I introduce a very important distinction. At the aggregate level, output is produced from capital and labor inputs. At the level of individual sectors, we find a role for capital and labor inputs, but also for inputs of energy and other intermediate goods. Rather than carrying over measures of output appropriate for economic aggregates to the sectoral level, we can define the value of output for each industrial sector to

include the value of capital, labor, energy, and other intermediate inputs. In Table 1, I have weighted the sectoral rates of technical change at the individual sectors by the total output of the sector, divided by the deliveries of output to final demand.

The other components that link technical change at the sectoral level to technical change at the aggregate level include the redistributions of output, capital input, and labor input among sectors. If we consider the period from 1960 to 1979, we see that the decomposition of the rate of technical change in Japan of 2.0 percent allocates 0.7 percent to rates of technical change at the sectoral level, 0.4 percent to the redistribution of outputs among sectors, 0.9 percent to the redistribution of capital input, and 0.1 percent to the redistribution of labor input.

In Table 1 we see that sectoral technical change accounts for only about a third of the aggregate technical change in Japan. The remaining two-thirds correspond to gains in efficiency that are not sustainable. These gains in efficiency result from the redistribution of the basic factors of production and the output of the different sectors. Redistributive gains are not sustainable since there is an upper limit to the amount of reallocation that can take place.

If we now consider the period from 1973 to 1979 we find that the weighted sum of sectoral rates of technical change in Japan went from a positive 0.7 percent to a negative 1.2 percent. In other words, the rate of technical change from 1973 to 1979 was negative in the average Japanese industry. In the United States, rates of technical change at the sectoral level declined from 0.4 to a negative 0.7 percent. This is a paradox, and it deserves an explanation.

In both Japan and the United States, production methods reverted to vintages of technological development that existed before the energy crisis—perhaps in the middle 1960's in Japan and the early 1960's in the United States. These earlier technological strata were appropriate to the new energy price situation. I conclude that it is perfectly consistent with a theory of economic growth to have negative rates of technical change, like the ones we see in Table 1 for the period 1973 to 1979.

<sup>2</sup>An excellent analysis of the slowdown in economic growth in industrialized countries is presented by Assar Lindbeck (1983). A leading proponent of the view that energy prices have no impact on economic growth is Edward Denison (1984).

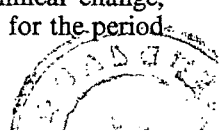


TABLE 2—CLASSIFICATION OF JAPANESE INDUSTRIES  
BY BIASES OF PRODUCTIVITY GROWTH

---

Capital Using, Labor Using, Energy Using, Material Saving
Agriculture, mining, construction, textiles, fabricated metal, transportation equipment, services
Capital Saving, Labor Using, Energy Saving, Material Saving
Machinery, finance and insurance
Capital Saving, Labor Using, Energy Using, Material Using
Food, petroleum
Capital Saving, Labor Using, Energy Using, Material Saving
Apparel, lumber, furniture, paper, printing, chemicals, rubber, leather, stone, clay, and glass, iron and steel, nonferrous metal, motor vehicles, instruments, miscellaneous manufacturing, transportation and communication, utilities, trade, real estate

---

Source: M. Kuroda, K. Yoshioka, and myself (1984).

We now have an even deeper mystery than before. The rate of technical change in Japan and the United States at the aggregate level is an unexplained residual and the same is true at the sectoral level. We find that dramatic changes in the rate of technical change at the sectoral level are behind the growth slowdown that we have observed. These changes are in the nature of unexplained residuals at the level of individual industries. At this point, I will use the concept of biased technical change to analyze the changes in economic growth that we have seen both in Japan and in the United States. In Table 2, I have classified industries in Japan by the pattern of biases of technical change. Four basic types of biases are related to capital, labor, energy, and materials inputs. We now require a more precise definition of the notion of a bias.<sup>3</sup>

The bias of technical change is the change in the relative share in the value of the output of a particular input as technology evolves. If we take energy as an example, we can say that if the share of energy in the value of the output of an industry is independent of the level of technology, then

technical change is unbiased or neutral with respect to energy. If the share of energy declines, we say that technical change is energy saving. If the share increases, technical change is energy using. We have a three-fold classification of technical change with respect to each input—capital, labor, energy, and materials—input using, input saving, and neutral.

One aspect of biased technical change relates to the direction of changes in the use of various inputs as technology evolves. For example, if a sector uses more capital, labor, and energy and saves materials, we have the pattern that is described in the first panel of Table 2. This pattern characterizes a substantial number of industries in Japan. However, there is a completely different implication of biased technical change. If technical change is energy using, then the rate of technical change declines when the price of energy increases. This provides the link between changes in energy prices and changes in the rate of technical change at the sectoral level.

Among all Japanese industries, only 3 out of the 30 listed in Table 2 are characterized by energy-saving technical change. In the other 27 industries, technical change is energy using. With unchanged input prices, the evolution of technology results in the use of more and more energy and a reduction in the use of the other inputs. The other implication of energy-using technical change is that if we have an increase in energy prices, there must be a corresponding reduction in the rate of technical change. In 27 out of the 30 Japanese industries, we have a direct link between energy prices and the rate of technical change through the energy using bias.

To make the link between sectoral rates of technical change more explicit, the typical Japanese industry described in Table 2 is characterized by energy-using technical change. This implies that when energy prices increase, as they did in 1973 and again in 1978, there will be a reduction in the rate of technical change in the average industry. We have already seen a decline in the weighted sum of sectoral rates of technical change for Japanese industries in the period 1973–79 in Table 1. I have identified this decline as the

<sup>3</sup>Biases of technical change are discussed by my papers (1983, 1984), and Kuroda et al.

TABLE 3—CLASSIFICATION OF U.S. INDUSTRIES  
BY BIASES OF PRODUCTIVITY GROWTH

---



---

Capital Using, Labor Using, Energy Using,
Material Saving
Agriculture, metal mining, crude petroleum and
natural gas, nonmetallic mining, textiles, apparel,
lumber, furniture, printing, leather, fabricated
metals, electrical machinery, motor vehicles,
instruments, miscellaneous manufacturing,
transportation, trade, finance, insurance and
real estate, services
Capital Using, Labor Using, Energy Saving,
Material Saving
Coal mining, tobacco manufactures,
communications, government enterprises
Capital Using, Labor Saving, Energy Saving,
Material Saving
Petroleum refining
Capital Using, Labor Saving, Energy Saving,
Material Using
Construction
Capital Saving, Labor Saving, Energy Using,
Material Saving
Electric utilities
Capital Saving, Labor Using, Energy Saving,
Material Saving
Primary metals
Capital Saving, Labor Using, Energy Using,
Material Saving
Paper, chemicals, rubber, stone, clay and glass,
machinery except electrical, transportation
equipment and ordnance, gas utilities
Capital Saving, Labor Saving, Energy Using,
Material Using
Food

---

Source: My paper (1983).

major explanatory factor in the slowdown of Japanese economic growth.

The weighted sum of sectoral rates at technical change dropped from 0.7 percent per year in Japan for the period 1960–79 to a negative 1.2 percent during the period 1973–79. The decrease of 1.9 percentage points more than accounts for the decline in the aggregate rate of technical change in Japan. This decline is the most important source of the slowdown in economic growth that occurred after 1973.

In Table 3, I consider the implications of biased technical change for the slowdown in U.S. economic growth. In the first panel of this table, observe that the character of technical change in the United States is predomi-

nantly capital using, labor using, energy using, and material saving. By contrast, the character of technical change in Japan is predominantly capital saving, labor using, energy using, and material saving. In both countries, technical change is characterized by using more energy. But in the United States, technical change also uses more capital as well as more labor, whereas in Japan, technical change uses less capital, less materials, and more labor along with more energy.

I have now arrived at the final explanation of the slowdown in U.S. and Japanese economic growth. I have emphasized that there are important sources of the slowdown in Japan associated with the falloff in upgrading of the labor force. However, the most important single factor in the Japanese slowdown is the sharp decline in the rate of technical change. I have now succeeded in linking that decline directly to energy prices through the energy using bias of technical change in Japan.

In the United States, the character of technical change is similar to that in Japan in the use of energy. The effect of higher energy prices in the United States was to slow economic growth. Of course, there are two additional facts that should be kept in mind. First, economic growth in the United States was a good deal less rapid than in Japan before the energy crisis. Second, energy prices increased much less substantially in the United States than in Japan. This is why the weighted sum of rates of technical change at the sectoral level decreased in the United States by only 1 percent, whereas in Japan, the decrease was nearly 2 percent.

My overall conclusion is that there was a dramatic impact of energy prices on economic growth during the energy crisis. The economic impact was very strongly negative in both the United States and Japan. The impact of higher energy prices was pervasive in the sense that it affected almost every industry in both economies. Almost every industry experienced a slowdown in the rate of technical change. This can be traced to the relationship between higher energy prices and the rate of technical change at the sectoral level in both countries.

## REFERENCES

- Christensen, Laurits R., Diane Cummings and Jorgenson, Dale W., "Economic Growth, 1947-1973: An International Comparison," in J. W. Kendrick and B. Vaccara, eds., *New Developments in Productivity Measurement and Analysis*, NBER Studies in Income and Wealth, 41, Chicago, University of Chicago Press, 1980, 595-698.
- \_\_\_\_\_, \_\_\_\_\_, and \_\_\_\_\_, "Relative Productivity Levels, 1947-1973," *European Economic Review*, May 1981, 16, 61-94.
- Denison, Edward F., "Accounting for Slower Economic Growth: An Update," in J. W. Kendrick, ed., *International Comparisons of Productivity and Causes of the Slowdown*, Cambridge: Ballinger, 1984, 1-45.
- Jorgenson, Dale W., "Modeling Production for General Equilibrium Analysis," *Scandinavian Journal of Economics*, No. 2, 1983, 85, 101-112.
- \_\_\_\_\_, "The Role of Energy in Productivity Growth," *Energy Journal*, July 1984, 5, 11-25.
- \_\_\_\_\_, Gollup, F. M. and Fraumeni, B. M., *Productivity and U.S. Economic Growth*, Cambridge: Harvard University Press, 1988.
- \_\_\_\_\_, Kuroda, Masahiro, and Nishimizu, Mieko, "Japan-U.S. Industry-Level Productivity Comparisons, 1960-1979," *Journal of the Japanese and International Economies*, March 1987, 1, 1-30.
- \_\_\_\_\_, and Mieko Nishimizu, "U.S. and Japanese Economic Growth, 1952-1974: An International Comparison," *Economic Journal*, December 1978, 88, 707-26.
- \_\_\_\_\_, et al., "Bilateral Models of Production for Japanese and U.S. Industries," in C. R. Hulten and J. R. Norsworthy, eds., *Productivity in the U.S. and Japan*, NBER Studies in Income and Wealth, Vol. 51, Chicago, University of Chicago Press, forthcoming 1988.
- Kuroda, Masahiro, Yoshioka, Kanji and Jorgenson, Dale W., "Relative Price Changes and Biases of Technical Change in Japan," *Economic Studies Quarterly*, 35, August 1984, 116-38.
- Lindbeck, Assar, "The Recent Slowdown of Productivity Growth," *Economic Journal*, March 1983, 93, 13-34.



# Industrial *R&D* in Japan and the United States: A Comparative Study

By EDWIN MANSFIELD\*

Given that Japan has become America's principal rival in many high-technology industries, it obviously is important from both an analytical and policy viewpoint to compare the extent, nature, organization, and effectiveness of the research and development (*R&D*) activities of Japanese firms with those of comparable American firms. This paper summarizes the results of a study, based both on published data and on data collected directly from a carefully selected sample of 200 Japanese and American firms, comparing the size and composition of firms' *R&D* expenditures in both countries, as well as the sources of their industrial *R&D* projects. In addition, I study the relationship between the extent of the industrial *R&D* in each country, on the one hand, and the rate of productivity increase, on the other hand, with particular emphasis on the differences between these countries in the effects of applied *R&D* and basic research.

## I. Industrial *R&D* and Productivity Growth

Economists are particularly interested in the relationship between industrial *R&D* and productivity growth. For the United States, there have been many studies of this topic, whereas for Japan there have been very few. This section presents the results of a study for Japan, the model and methods used being

similar to those employed in earlier studies by Griliches, Terleckyj, and myself. (See my 1980 article.) In a particular industry, the production function is assumed to be

$$(1) \quad Q_t = Ae^{\lambda t} R_t^\alpha L_t^\gamma K_t^{1-\alpha-\gamma},$$

where  $Q_t$  is the industry's value-added in year  $t$ ,  $R_t$  is the industry's stock of *R&D* capital,  $L_t$  is the industry's labor input, and  $K_t$  is the industry's capital input. Thus, the annual rate of change of total factor productivity is

$$(2) \quad \rho_t = \lambda + \Phi((dR/dt)/Q_t),$$

where  $\Phi = \partial Q_t / \partial R_t$ .

Because the rate of growth of each industry's stock of *R&D* capital cannot be measured directly, I assume, as other studies frequently have, that an industry's expenditure on *R&D* during year  $t$  is equal to the change in the industry's stock of *R&D* capital then. If an industry's values  $\lambda$  and  $\Phi$  are statistically independent of its ratio of *R&D* expenditure to value-added, it follows that

$$(3) \quad \rho_i = \bar{\lambda} + \bar{\Phi}r_i + z_i,$$

where  $\rho_i$  is the annual rate of change of total factor productivity in the  $i$ th industry,  $r_i$  is the ratio of *R&D* expenditure to value-added in this industry,  $\bar{\lambda}$  and  $\bar{\Phi}$  are the average values of  $\lambda$  and  $\Phi$  in all relevant industries, and  $z_i$  is a random error term.<sup>1</sup>

\*Director, Center for Economics and Technology, University of Pennsylvania, Philadelphia, PA 19104. The research on which this paper is based was supported by a grant from the Division of Policy Research and Analysis of the National Science Foundation. I am indebted to Rolf Piekarz, Gary Saxonhouse, and Eleanor Thomas for helpful comments, as well as to the 200 Japanese and American firms that provided much of the basic data. I also thank Masahiro Kuroda, Keio University, as well as the OECD, for providing me with unpublished data. A more complete version of this paper describing the analysis and results in more detail is available from the author.

<sup>1</sup>The only previous study (that I know of) using industry-level data to estimate the relationship between *R&D* and productivity growth in Japan is Hiroyuki Odagiri (1985). Zvi Griliches and Jacques Mairesse (1985) have used firm-level data, and Pierre Mohnen, M. Ishaq Nadiri, and Ingmar Prucha (1985) have used manufacturing sector data for this purpose. There have been no attempts to disaggregate *R&D*, the principal focus of attention here.

TABLE 1—REGRESSION COEFFICIENTS TO EXPLAIN  
RATES OF PRODUCTIVITY INCREASE,  
2-DIGIT MANUFACTURING INDUSTRIES,  
JAPAN AND THE UNITED STATES

$r_i$	Independent Variables <sup>a</sup>			$R^2$
	$I_i$	$A_i$	$B_i$	
Japan (1960–79) <sup>b</sup>				
0.42 (2.78)	—	—	—	0.34
0.33 (1.88)	0.05 (0.91)	—	—	0.38
—	0.05 (0.95)	0.54 (1.59)	—1.52 (0.58)	0.40
—	—	0.60 (1.84)	—1.23 (0.48)	0.36
United States (1948–66) <sup>c</sup>				
—	—	0.07 (2.75)	1.49 (3.34)	0.63

<sup>a</sup> $I_i$  is the proportion of technology import agreements in the  $i$ th industry,  $A_i$  is the ratio of applied R&D expenditure to value-added in the  $i$ th industry, and  $B_i$  is the ratio of basic research expenditure to value-added in the  $i$ th industry. The number in parentheses below each regression coefficient is its  $t$ -value.

<sup>b</sup>The Japanese industries included are the same as those used in Odagiri, except that (i) textiles and apparel, and (ii) motor vehicles and other transportation equipment are not combined. Thus, the sample size is 17.

<sup>c</sup>Another variable included in the U.S. regression is the percent of an industry's workers that are unionized. Many investigators have found this variable to be significant in the United States, but Odagiri reports that it has not been significant in Japan. Frequently, in models of this sort, the amount of R&D that is embodied in an industry's purchased inputs (divided by its value-added) is used as another explanatory variable. My article (1980) found that it was significant in the United States, but Odagiri's results suggest that it is not significant in Japan.

To estimate this model, I used Kuroda's data concerning  $\rho_i$  for 2-digit manufacturing industries in Japan for 1960–79 and his data concerning 1972 value-added in these industries, together with OECD data concerning 1972 R&D expenditures, to calculate  $r_i$ . The least squares estimate of  $\bar{\Phi}$ , shown in Table 1, indicates that, in Japan, an industry's rate of productivity growth was significantly related to its ratio of R&D expenditure to value-added during this period. Since previous studies often have used estimates of  $\bar{\Phi}$  as measures of the rate of return from R&D, it is interesting to note that, if such an interpretation is adopted, the rate of return was about 42 percent, which is relatively high. However, as many of us have stressed, such measures of the rate of return

from R&D are exceedingly crude, particularly for industry-level data.<sup>2</sup>

Because R&D-intensive industries have tended to be the biggest importers of foreign technology, some of the apparent effect of domestic R&D on productivity increase may really have been the effect of imported technology. (As is well known, Japan's relatively rapid rate of technological change has been due largely to the importation of foreign technology.) To see whether this is the case, I introduced as an additional variable in equation (3) the proportion of technology import agreements during 1950–65 in the  $i$ th industry. As expected, the regression coefficient of  $r_i$  falls when this variable is introduced, but not by much. It still is about 33 percent, which is relatively high.

## II. Basic Research vs. Applied R&D

In the United States, the variation among industries (and firms) in  $\rho_i$  can be explained more completely if basic research expenditure is distinguished from applied research and development expenditure in equation (3). To see whether this was true in Japan as well, I used two variables—the ratio of applied R&D expenditure to value-added and the ratio of basic research expenditure to value-added—in place of  $r_i$  in equation (3). As shown in Table 1, the ratio of basic research to value-added does not have a statistically significant effect on  $\rho_i$ . However, the estimated regression coefficient of the ratio of applied research and development to value-added is about 60 percent, which is quite large.

It is interesting to compare these regression results for Japan with my earlier results for the United States (reproduced in part in

<sup>2</sup>The effects of R&D often occur with a lag (which for basic research can be quite long). Because so little is known about these lags, it is very difficult to include them properly in a model of this sort (see my 1980 article). Also, a variety of other problems and limitations have been identified, many of which are less serious for firm-level than industry-level data. Despite these criticisms, these measures have been used repeatedly to estimate rates of return.

Table 1). If we ignore the many reasons why these regression coefficients are likely to be very crude measures of rates of return, such a comparison suggests that American firms have obtained higher returns from basic research than the Japanese, whereas the Japanese have obtained higher returns from applied *R&D* than the Americans. Such a result seems reasonable. Because Japan has been able to draw at relatively small cost on a rich stock of foreign technology that was more advanced than its own, and because a relatively small percentage of Japan's industrial *R&D* has been financed by the government (and has gone largely for non-commercial purposes), it is entirely reasonable that the rate of return from applied *R&D* may have been higher in Japan than in the United States. Also, Japan's emphasis on process rather than product technology (discussed in Section III) may have enhanced the payoff from its applied *R&D*. Because of differences between the two countries in the extent of the external benefits to industrial basic research from university research, it is also reasonable that the rate of return from industrial basic research may have been lower in Japan than in the United States. In the United States, there often have been close working relationships between basic researchers in industry and their colleagues in the universities. In Japan, university research seems to have played a lesser role (and seems to have been less highly regarded) than in the United States.

### III. Industrial *R&D*: Intensity and Composition

If it is true that the rate of return from applied *R&D* in Japan has been relatively high, one can readily understand why the *R&D* intensity of manufacturing firms has increased more rapidly in Japan than in the United States. In 1986, company-financed *R&D* expenditures were about 2.7 percent of sales in Japan, in comparison with about 2.8 percent in 1985 in the United States. In 1970, the corresponding figures were 1.3 percent for Japan and 2.2 percent for the United States. In all industries other than machinery, instruments, paper, and petroleum, Japan has narrowed the gap substan-

tially. In some industries (food, textiles, metals, and rubber) Japan now leads; in other industries (paper, petroleum, machinery, and instruments) the United States now leads; and in the rest there is a relatively small difference in *R&D* intensity.

Since data on total *R&D* expenditures, while useful, are difficult to interpret because *R&D* projects are so heterogeneous, I collected data concerning the composition of their *R&D* expenditures from a carefully selected sample of Japanese and American firms. Fifty Japanese firms were chosen at random in the chemical, electrical equipment, instrument, machinery, rubber, and metals industries, and for each Japanese firm we picked at random an American firm of the same industry and approximate size. The firms in our sample carry out about 25 percent of the *R&D* in each country in these industries. Detailed information was procured from each of these 100 firms (50 matched pairs) concerning the percentage of its 1985 *R&D* expenditures devoted to 1) basic research, 2) applied research, 3) product (rather than process) technology, 4) projects aimed at entirely new products and processes, 5) projects with less than a 0.5 estimated chance of success, and 6) projects expected to last longer than 5 years.

Based on the results, the Japanese seem to devote about as large a percentage of their *R&D* expenditures to relatively risky and long-term projects as do American firms (Table 2). This differs greatly from the early 1970's, when Merton Peck and Shuji Tamura (1976) characterized Japanese industrial *R&D* as composed very largely of "low-risk and short-term projects." Nonetheless, it would be a mistake to think that Japanese and American industrial *R&D* have become essentially the same. Whereas American firms report that almost half of their *R&D* expenditures are going for projects aimed at entirely new products and processes, Japanese firms report that only about one-third of their *R&D* expenditures go for this purpose. (Outside the chemical industry, where there is little difference in this regard, the gap is even wider.) Of course, this is in accord with a great deal of anecdotal information to the effect that the Japanese devote

TABLE 2—COMPOSITION OF *R&D* EXPENDITURES,  
100 FIRMS (50 MATCHED PAIRS),  
JAPAN AND THE UNITED STATES, 1985<sup>a</sup>

Percent of <i>R&amp;D</i> Expenditures Devoted to:	Japan	United States
Basic Research	10	8
Applied Research	27	23
Products (rather than processes)	36	68
Entirely New Products and Processes	32	47
Projects with less than 0.5 Estimated Chance of Success	26	28
Projects Expected to Last Longer than 5 Years	38	38

<sup>a</sup>The number of firms in each industry is chemicals (including drugs), 36; electrical and instruments, 20; machinery (including computers), 30; and rubber and metals, 14.

more of their *R&D* resources to the improvement and adaptation of existing products and processes (rather than to the development of entirely new products and processes) than do American firms.

Even more striking is the difference between Japanese and American firms in their allocation of *R&D* resources between projects aimed at improved *product* technology and projects aimed at improved *process* technology. The American firms in my sample devote about two-thirds of their *R&D* expenditures to improved product technology (new products and product changes) and about one-third to improved process technology (new processes and process changes). Among the Japanese firms, on the other hand, the proportions are reversed, two-thirds going for improved process technology and one-third going for improved product technology. Harking back to my results in Section II, Japan's greater emphasis on process technology probably accounts in part for its relatively high estimated value of  $\phi$ , since process *R&D* tends to have a bigger effect on an industry's own rate of productivity increase than does product *R&D*.<sup>3</sup>

<sup>3</sup>From the point of view of the economy as a whole, a large proportion of the resources allocated to product technology in the United States really goes for processes,

These results shed new light on a major issue concerning industrial *R&D* in the United States. Many observers have criticized American industry for neglecting process innovation. As the President's Commission on Industrial Competitiveness puts it, "It does us little good to design state-of-the-art products, if within a short time our foreign competitors can manufacture them more cheaply" (1985, p. 20). Contrary to the common impression that U.S. firms have in recent years begun to react to such criticism by paying more attention to process innovation than in the past, my results do not indicate that there was any perceptible increase between 1976 and 1985 in the proportion of their *R&D* expenditures devoted to new or improved processes. Thus, in terms of the allocation of their *R&D* funds, American firms do not seem to have put more emphasis on processes, despite this criticism.

#### IV. Industrial *R&D*: Sources and Size of Firm

Some important differences between the industrial *R&D* efforts in Japan and the United States can be highlighted by looking at the sources of *R&D* projects in the two countries. Detailed data on this score were obtained from a random sample of 65 American and 35 Japanese firms in the chemical, electrical equipment, machinery, motor vehicle, instruments, and metals industries.<sup>4</sup> The results (in Table 3) indicate

since one firm's products frequently are parts of another firm's processes. Consequently, this difference between Japan and the United States reflects a difference in how much of the process *R&D* for a given product is carried out by the producers of the product and how much is done by equipment producers and other suppliers of the producers of the product. As many authors have stressed, there can be disadvantages in leaving this sort of *R&D* to the latter firms. (Also, differences in industry and firm structure are relevant.) Many studies indicate that process *R&D* has a bigger effect on productivity than does product *R&D*. However, many of the effects on productivity resulting from new products occur in the industries using them, not in those selling them, and there are difficulties in taking proper account of quality changes, as well as other problems in measuring the benefits from new products.

<sup>4</sup>The American firms were chosen at random from the listing of major firms in these industries in *Business Week*, July 9, 1984. The Japanese firms were chosen at

TABLE 3—SOURCES OF R&amp;D PROJECTS, 100 FIRMS, JAPAN AND UNITED STATES, 1985

Industry/ Country <sup>a</sup>	Percent of R&D Projects Suggested by:			
	R&D	Marketing	Production	Customers
Total				
Japan	47	18	15	15
U.S.	58	21	9	9
Chemicals				
Japan	49	23	15	3
U.S.	45	25	14	8
Electrical				
Japan	47	21	5	27
U.S.	90	7	1	1
Machinery				
Japan	44	22	11	20
U.S.	56	21	4	18
Autos, Instruments, and Metals				
Japan	48	8	26	13
U.S.	51	25	12	11

<sup>a</sup>The sample sizes are all industries combined, 100; chemicals, 26; electrical, 20; machinery, 26; and autos, instruments, and metals, 28.

that Japanese firms base about one-third of their R&D projects on suggestions from their production personnel and customers, whereas only about one-sixth of American projects stem from these sources. The greater importance of production personnel as sources of R&D projects in Japan is a reflection of their greater emphasis (described in Section III) on process technology. The greater importance of customers as sources of R&D projects in Japan stems from the very close relations there, noted by Henry Riggs (1985) and others, between firms and their customers. What is especially noteworthy is that both production personnel and customers tend to be users of a firm's R&D results, and that the Japanese seem to give users a more important role than do Americans in shaping their R&D programs.

Particularly in the electrical equipment industry, American firms tend to base a larger percentage of their R&D projects on suggestions from R&D personnel than do Japanese firms. This is in accord with Riggs' hypothesis that, whereas Japanese firms are better able to carry out advances inspired or driven by users, American firms are better able than their Japanese rivals "to capitalize on opportunities...which are derived from, or in-

spired by, technology (the research laboratory)..." (p. 10). Also, it is consistent with Masohiko Aoki's belief that American firms put more emphasis than do Japanese firms on the "scientific efforts of professionally trained researchers in R&D under the entrepreneurial direction of top management" (1986, p. 981).

Another difference between Japan and the United States pertains to R&D expenditures aimed at entirely new products and processes. In the United States, increases in firm size tend to be associated with less than proportionate increases in the amount spent on such R&D. In Japan, on the other hand, increases in firm size tend to be associated with more than proportionate increases in the amount spent on such R&D. Thus, relative to smaller firms, the largest firms tend to carry out a disproportionately greater amount of this very ambitious R&D in Japan than in the United States. This may be one reason why small and moderate-sized firms in Japan seem to have contributed less to innovation than their counterparts in the United States.

## V. Conclusions

Many observers are impressed by the efficiency of Japanese industrial R&D. Indeed, the president of the Semiconductor Research Corporation has gone so far as to state that: "The United States may never match Japan's R&D efficiency." (See L. Sumney and R. Burger 1987, p. 40.) If one is willing to interpret the regression coefficients in Table 1 as rates of return, my results are consistent with the contention that applied R&D in Japan has yielded a higher return than in the United States. This contention seems reasonable, given Japan's greater emphasis on commercial (rather than government-financed) projects and its reliance on advanced technology from the West, which could be adapted and improved at relatively low cost. But this is only part of the story. My findings provide for the first time data showing the great extent to which Japan has focused on process technology, which according to many experts has tended to be neglected in the United States. Also, I show elsewhere (forthcoming) that Japanese firms

random from *International Dun and Bradstreet*. These same frames were used for the samples in Table 2.

seem to have been much faster and more efficient imitators than American firms. Almost certainly, these factors too have contributed to the effectiveness of applied *R&D* in Japan.

On the other hand, there is no evidence that basic research has been relatively fruitful in Japan. Moreover, based on the findings in my forthcoming paper, there is no evidence that Japanese firms have been faster or more efficient innovators than American firms in cases where the innovation has been based on internal, rather than external, technology. Apparently, the Japanese advantage has been confined largely to applied *R&D*, particularly *R&D* concerned with the adaptation and improvement of existing technology.

Faced with the Japanese technological challenge, American firms might respond by putting more resources into process *R&D*, which would make it more difficult for Japanese firms and others to appropriate a large share of the benefits from American product innovations. Also, American firms might increase their own capacity to imitate quickly, efficiently, and creatively. If they respond effectively, there is no reason why the United States cannot increase the economic returns from its industrial *R&D*, although it is inevitable—and by no means undesirable, both from the point of view of the United States and of the world as a whole—that many of the economic benefits from this *R&D* will continue to accrue to other nations.

## REFERENCES

- Aoki, Masohiko, "Horizontal vs. Vertical Information Structure of the Firm," *American Economic Review*, December 1986, 76, 971–83.
- Griliches, Zvi and Mairesse, Jacques, "*R&D* and Productivity Growth: Comparing Japanese and U.S. Manufacturing Firms," NBER Working Paper, 1985.
- Jorgenson, D., Kuroda, M. and Nishimizu, M., "Japan-U.S. Industry-Level Productivity Comparison, 1960–79," Conference on Research in Income and Wealth, 1985.
- Mansfield, Edwin, "The Speed and Cost of Industrial Innovation in Japan and the United States: External vs. Internal Technology," *Management Science*, forthcoming.
- \_\_\_\_\_, (1987a) "The Diffusion of Industrial Robots in Japan and the United States," Center for Economics and Technology, University of Pennsylvania, 1987.
- \_\_\_\_\_, (1987b) "Firm Growth, Innovation, and *R* and *D* in Robotics: Japan and the United States," Symposium on Research and Development, Industrial Change, and Economic Policy, University of Karlstad, Sweden, 1987.
- \_\_\_\_\_, "Basic Research and Productivity Increase in Manufacturing," *American Economic Review*, December 1980, 70, 863–73.
- Mohnen, Pierre, Nadiri, M. I. and Prucha, Ingmar, "*R* and *D*, Production Structure, and Rates of Return in the U.S., Japanese, and German Manufacturing Sectors," *European Economic Review*, 1986, 30, 749–71.
- Odagiri, Hiroyuki, "Research Activity, Output Growth, and Productivity Increase in Japanese Manufacturing Industries," *Research Policy*, 1985, 14, 117–30.
- Peck, Merton and Tamura, Shuji, "Technology," in H. Patrick and H. Rosovsky, eds., *Asia's New Giant*, Washington: The Brookings Institution, 1976.
- Riggs, Henry, "Innovation: A U.S.–Japan Perspective," Stanford University High-Technology Research Project, September 1985.
- Sumney, L. and Burger, R., "A Semiconductor Strategy," *Issues in Science and Technology*, Summer 1987, 3, 32–41.
- President's Commission on Industrial Competitiveness, *Global Competition: The New Reality*, Washington: USGPO, 1985.

# Why are Americans Such Poor Imitators?

By NATHAN ROSENBERG AND W. EDWARD STEINMUELLER\*

Despite American success in previous historical eras in imitating the technology and organizational structure of our industrial rivals in other nations, there is mounting evidence that our capacity to absorb and adapt our rivals' advantages to our own purposes has diminished in recent years. While these concerns are voiced with regard to a number of nations, the recent success of Japanese firms has been noteworthy and deserves special attention.

One reason Americans have been such poor imitators is that, until very recently, we were not even aware that there was much in Japanese industry that was worth imitating. Japanese economic competitiveness was, for a long time, dismissed as simply reflecting lower labor costs, which were regarded as decisive in certain industries. Later, Japanese success was dismissed as ephemeral, reflecting the ease of rapid growth on the part of a "mere imitator" following the innovative leads of other nations, particularly our own.

More recently, as competition has become more heated and as certain American industries have suffered heavily from Japanese imports, the successes of Japanese firms have been attributed to policies of "industrial targeting" orchestrated by the Ministry of International Trade and Industry (MITI), usually said to involve extensive government subsidies and coordination of import policies that unfairly tilted what should have been a level playing field.

We do not wish to deny that there may have been some truth to each of these beliefs at one point in time. However, an unfortunate consequence of such beliefs has been that they have delayed efforts to monitor

and study the performance of the Japanese manufacturing sector with any care.

Certainly an earlier complacency has now unravelled. It is abundantly clear that there is much to admire, and perhaps to emulate, in some parts of the Japanese manufacturing system, particularly in the production of goods with a high degree of systemic complexity—for example, plain paper copiers, automobiles, some machine tools, and consumer electronics. In retrospect it is obvious that there has been much that has been worth imitating, but Americans, even when they have become aware of this, have been poor imitators. Why should this have been so? Why have the Japanese been so much better at imitation than the Americans? What has made the Japanese, if we may be permitted to use the phrase, such "creative imitators?"

During the past twenty years, the composition of Japanese exports to the United States has shifted dramatically away from industries where labor intensity provided comparative advantage, to industries where sophisticated manufacturing skills and technology are of central importance. This shift has been so marked that, in 1985, over 80 percent of Japanese exports to the United States were in electrical, electronic, transportation equipment, and machinery sectors. In short, Japanese "imitation" has been concentrated in a few specific sectors where American industry had previously been dominant for many years following World War II. The imitative processes that were used by Japanese industry in catching up to U.S. firms in these sectors should now be recognized as having major implications for how Japanese firms are prepared to succeed at the more difficult task of forging ahead (Moses Abramovitz, 1986).

The first part of our answer to the question of why Americans have been such poor imitators is that there has been a distinct asymmetry in the strengths developed by

\*Fairleigh S. Dickinson, Jr. Professor of Public Policy, Department of Economics, Stanford University, Stanford, CA 94305, and Deputy Director, Center for Economic Policy Research, 100 Encina Commons, Stanford University, Stanford, CA, 94305, respectively.

each of these industrial economies. This asymmetry partly accounts for why it has taken so long to appreciate fully the sources of Japanese industrial capabilities. The Japanese have been very successful in borrowing and developing technologies initially created by American firms. These technologies have been largely of a hardware nature, in particular, a stream of highly visible product innovations.

By contrast, what may be most worth imitating on the Japanese side is much more subtle and much less visible. It includes ways in which certain *activities* are carried out, rather than readily identifiable pieces of hardware. These differences lie at the levels of organization and incentives for improvement. The first is the efficient coordination of product design and manufacturing functions. The second is effective solutions to the myriad small problems that are key to efficient mass production techniques. An important part of the reason that it has been so difficult to appreciate the nature of these Japanese achievements is that they are heavily concentrated in a collection of activities—development—that economists have, so far, failed to unpack and subject to detailed and critical analysis.

This is surprising for a number of reasons, not the least of which is that *R&D* is, in fact, overwhelmingly *D*. Yet, we know more about the 12 percent of *R&D* that constitutes basic research than of the 68 percent that constitutes development. While this may be understandable on the part of natural scientists, it is less so on the part of economists. Nevertheless, American thinking about the innovation process has focused excessively upon the earliest stages—the kinds of new products or technologies that occasionally emerge out of basic research, the creative leaps that sometimes establish entirely new product lines, the activities of the “upstream” inventor or scientist rather than the “downstream” engineer. American discussions of technical change are more likely to be presented in terms of major innovations and pioneering firms, rather than in terms of the success of particular sectors or firms at catching up and overtaking other organizations through sustained effort and small im-

provements. In this respect, the dominant view of the innovative process is still overly Schumpeterian, in its preoccupation with discontinuities and creative destruction, and its neglect of the cumulative power of numerous small, incremental changes. We suggest that the Japanese have had a much deeper appreciation of the economic significance of these vital development activities than their American counterparts.

Development, of course, covers a range of activities whose content differs widely from one industry to another. It generally includes the designing of new products, testing and evaluating their performance (which in some industries may involve the building and testing of prototypes, or experimentation with pilot plants), and inventing and designing new and appropriate manufacturing processes. In each of these activities, the role of minor modifications and small improvements that better integrate design and production, establish closer feedbacks from users to suppliers, and more effectively “tune” existing production methods, are critically important. Individually, each of these modifications and improvements will bring about some slight reduction in cost or improvement in performance. Their cumulative effects may, however, be immense, as when the semiconductor industry moves, through a multitude of small steps, from a handful of transistors on a chip to a million such transistors, or when the channel capacity of a 3/8” coaxial cable expands, through a succession of small improvements, by more than an order of magnitude, or when the speed of computers increases by several orders of magnitude.

It is the essence of these development activities that they have no well-defined terminus. They do not end when a new or improved product is brought to market. Quite the contrary. A continual stream of small improvements is often the essence of success in the competitive process. In industries such as those that currently account for the bulk of Japanese exports to the U.S., development is a never-ending activity. They are not, from some points of view, very exciting activities. They are activities that do not win Nobel Prizes; nor, for the most part,



do they even win recognition at the Patent Office. This low visibility accounts for the very limited awareness of their economic importance. Nevertheless, poor performance in the development process can be commercially fatal to firms that are highly successful at research. Such poor performance can readily translate into final products of inferior design, lower quality, and poor reliability. It can also translate into higher cost and, therefore, inability to sustain a market position originally achieved through the innovation process. These shortfalls can convert a technological head start, resulting from successful innovation, into a scramble to retain what turns out to be a shrinking market share against the cost and performance advantages of competitors, including those who may have had no role in the initial innovation or in the antecedent research that made it possible.

These possibilities are, of course, not offered as mere idle rumination. There is an accumulation of evidence that many Japanese successes in recent years are a consequence of greater effectiveness in organizing and providing strong incentives for these "downstream" development activities. In the internal organization of their firms, the Japanese commonly provide for much closer interaction between product designers and production engineers, they devote far more attention to the refinement of the appropriate process technologies, and they also assign a more prominent role to the engineering department.<sup>1</sup> In considerable measure, then, their skill in imitation has been an accompaniment of their skill in, and concern with, development activities. The significance of these activities is heightened by a recognition that the ability to imitate and

improve upon one's own prior performance, rather than starting from scratch, is often central to success at development activities. If American industry were to improve its development skills it would also, simultaneously, improve its capacity to imitate. The two capabilities overlap heavily.

This statement applies with particular force to the cultivation of a strong interface between product design and engineering. Japanese strength at this interface has facilitated its technology-imitation activities by the ease with which it enables foreign products to be quickly adapted and modified to suit domestic requirements, and low production costs to be speedily achieved (see Mansfield). Furthermore, it has made it possible to move to positions of leadership where new technologies call for simultaneous optimization on both the process and product sides. The more rapid exploitation of robotics in Japan appears to be due, in important measure, to the alacrity with which Japanese firms modified and simplified product design in order to accommodate the new robotics technology. It has probably been more sensible to simplify the design of products so that robots could readily assemble them—reducing the number of component parts and simplifying the method by which parts are attached to one another—than to design robots of more general, and therefore more sophisticated, assembling capabilities.

A central theme in the study of the development process has been its integrated, interactive, and iterative nature. In sharp contrast, American firms have often compartmentalized research and manufacturing functions. Often, this has led to breakdowns in the development process characterized by "finger pointing," in which functionally specialized groups within the firm assign blame to each other or to external suppliers. In spite of this, U.S. firms are often very good at innovation since individual ingenuity and sharply focused specialization can overcome many obstacles. But these same firms often find it difficult to make the small steps that are crucial to the ongoing development process. This leaves competitors with a host of opportunities for imitation and modification for improving performance or reducing costs,

<sup>1</sup>In a recent comparison of innovation in Japan and the United States, Edwin Mansfield (1988) has observed a striking difference with respect to the allocation of R&D budgets between product and process technology. According to Mansfield, the American firms in his sample devoted 2/3 of their R&D budgets to improved product technologies and only 1/3 to improved process technologies, whereas among the Japanese firms only 1/3 of the R&D budgets were devoted to improved product technologies and 2/3 to improved process technologies.

thereby truncating the appropriation of returns from innovation.

The Japanese have, on numerous occasions, been the leaders in the commercialization of new products, in spite of the fact that the new product, or some essential component, was invented elsewhere. Although the United States pioneered both the scientific and technological frontiers in the invention of the transistor, Japanese firms were the first to succeed in large-scale application of this technology for radios, and later obliterated America's earlier dominance of the market for color television receivers. Japanese success at quality and design improvements for mass-produced goods such as compact automobiles and consumer electronics are highly visible. Products requiring smooth coordination of different technologies (for example, electrical, electronic, and mechanical) for such things as plain paper copiers, facsimile machines, floppy disk drives, and personal computer printers, are strongholds of Japanese commercial and export success. None of these technologies rests on a single critical innovation. Instead, Japanese success in each of these areas can be traced to the cumulative impact of its great development capabilities.

Japanese success in development has often been able to overcome America's much-heralded innovative capabilities. The more specialized an activity becomes, the greater the importance of efficient information exchanges if inappropriate tradeoffs or inappropriate optimization criteria are to be avoided. For specialists to work well in a large organization, there must be an intimate familiarity with one another's goals and priorities. There must be a set of shared understandings and concerns. The development efforts of Japanese firms strongly emphasize rotation of personnel among departments in ways that lead to the exchange of useful information and the formation of common goals. In many cases, close communication among functionally separate specialists is strengthened by the awareness of a commonality of interest flowing from stable, long-term employment (and supplier) relationships. Japanese firms appear to make more systematic use of engineering skills and

production worker experience throughout the entire sequence of development activities associated with the introduction of new products, including the most minute aspects of the eventual manufacturing process.

These activities are not well appreciated when, as is commonly the case, development is thought of as the *application* of scientific knowledge.<sup>2</sup> Development in fact incorporates knowledge from many sources. Even in those instances in which new scientific knowledge *does* provide the initial stimulus for a new product, the subsequent development process will draw upon a wide variety of sources, the most common of which is likely to be the existing "in-house" engineering knowledge. Organizational structures and incentive systems that can exploit these sources effectively will create economic advantages over competitors who cannot do so, even if these competitors have superior research capability. If these development capabilities are sufficiently strong, the stage of commercialization may be reached sooner, and will certainly be reached by firms in a better position to subsequently reduce cost and improve performance (Masahiko Aoki and Rosenberg, 1987).

In short, the economic value of "first-mover" advantages in capturing the economic returns from innovation is overrated, because innovations are commonly very poorly designed in their earliest stages and in numerous ways ill-adapted to their ultimate applications (Rosenberg, 1976, ch. 11). The incremental improvements underlying development play a critical role in the eventual capture of returns from innovation.

Thus, there are two reasons for the primacy of development in capturing the returns from innovations in markets such as those in which the Japanese have demonstrated success. The first is that efficiency gains in mass production are often easier to achieve through large numbers of small improvements than

<sup>2</sup> The National Science Foundation defines Development as "...the systematic use of knowledge or understanding *gained from research* directed toward the production of materials, devices systems, or methods, including design and development of prototypes or processes" (1985, p. 221, emphasis added).

through major revisions. The second is that cost reduction and performance improvements in a well-established technology are often capable of overtaking efforts to advance technology through discontinuous "leaps" or major innovative steps. The creative elements of imitation involve not only the adaptation of new or externally created technology, but a continuing refinement of existing technologies and manufacturing methods.

The expansion of Japanese industries in the sectors accounting for most of their exports to the United States is largely the consequence of success at development activities (see our 1988 paper). In the consumer electronics industry, successful Japanese development of video cassette recording (VCR) technology employed multiple prototype development efforts, close coordination of design and manufacturing, and numerous small improvements in a product of considerable systemic complexity (Richard Rosenbloom and Michael Cusanamo, 1987). In addition to providing Japanese firms with a major export market, continuing improvements in VCR products well after their initial introduction were a major factor in RCA's failure to gain a market for video player technology (Margaret Graham, 1986). The same sorts of development activities, involving numerous small improvements and precise coordination of design and manufacturing, have been important in the commercialization of laser technologies for compact disc players and laser printers, markets that are currently experiencing major expansion.

The role of sustained incremental improvement and focus on manufacturing processes have also been important in electronics industries located upstream from the production of systems for end users. The technological innovativeness of the integrated circuit (IC) industry at producing intermediate goods for other electronics companies is by now well known (Ernest Braun and Stuart MacDonald, 1982). What has been less well appreciated by economists is that innovations in the IC industry have been strongly influenced by the incremental improvement of process technology (Steinmueller, 1987). Recent successes of Japanese

IC firms in international competition have been heavily dependent upon success at manufacturing improvement. In IC production, the proportion of workable devices emerging from the production process, production yield, is the most important manufacturing cost factor (Steinmueller, 1987), and yield is very sensitive to both the extent of production experience and its successful integration with the design process. Japanese accomplishments in international IC competition involves several factors (Steinmueller, forthcoming), including successful product and process development and high yields in the large-scale production of IC memory devices. The implications of development success have not been confined within the Japanese IC Industry. The demands of Japanese consumer electronics provided important impetus to the development of CMOS (complementary metal oxide semiconductor) technology. In part due to continuing development efforts, CMOS recently has emerged as the leading technology for future very large scale IC devices. As a consequence, several international joint ventures and other agreements are creating a flow of technological knowledge from Japanese to American IC firms (Steinmueller, forthcoming 1988).

Close communication links between suppliers and users play a role at the interfirm level that is analogous to our emphasis on effective communication links among functional specialists at the intrafirm level. Recent detailed studies of the organization of parts purchases in the Japanese automobile industry by Banri Asanuma (1985) demonstrate the existence of long-term relationships with important institutional mechanisms for coordinating design and assuring timely supply. Aoki (1987, p. 335) cites an (unnamed) major auto manufacturer as having 122 stable "first tier" suppliers. More importantly, Aoki characterizes these relationships as "quasi-permanent," noting that between 1973 and 1984 only 3 firms exited from this relationship while 21 firms entered. The consequences of such stable supplier relationships are that development efforts can be jointly initiated and pressed forward, further extending the coordination of product

design and manufacturing beyond the level of the individual firms as well as improving the flow of information for making modifications and improvements in the manufacturing process.

We draw an ironic conclusion from our examination of American and Japanese technological skills. The Japanese have indeed been excellent imitators. But instead of flourishing a trump card stating that Americans are excellent *innovators*, we need to fix our attention on the disconcerting prospect that innovative skills may count for a great deal less than we once thought—unless we can learn to become better imitators ourselves.

## REFERENCES

- Abramovitz, Moses, "Catching Up, Forging Ahead, and Falling Behind," *Journal of Economic History*, June, 1986, 46, 385-406.
- Aoki, Masahiko, "A Microtheory of the Japanese Economy: Information, Incentives and Bargaining," Kyoto Institute of Economic Research, Discussion Paper No. 241, Kyoto University, October, 1987.
- and Rosenberg, Nathan, "The Japanese Firm as an Innovating Institution," Center for Economic Policy Research, Publ. No. 106, Stanford University, September, 1987.
- Asanuma, Banri, "The Organization of Parts Purchases in the Japanese Automotive Industry" and "The Contractual Framework for Parts Supply in the Japanese Automotive Industry," *Japanese Economic Studies*, Summer 1985, 13, 32-78.
- Braun, Ernest and MacDonalds Stuart, *Revolution in Miniature: The History and Impact of Semiconductor Electronics*, 2d ed., Cambridge: Cambridge University Press, 1982.
- Graham, Margaret B. W., *RCA and the VideoDisc: The Business of Research*, Cambridge: Cambridge University Press, 1986.
- Mansfield, Edwin, "Industrial R&D in Japan and the United States: A Comparative Study" *American Economic Review Proceedings*, May 1988, 78, 223-28.
- Rosenberg, Nathan, *Perspectives on Technology*, Cambridge: Cambridge University Press, 1976.
- and Steinmueller, W. Edward, "Can Americans Become Better Imitators?," Center for Economic Policy Research, Publ. No. 117, Stanford University, 1988.
- Rosenbloom, Richard S. and Cusumano, Michael A., "Technological Pioneering and Competitive Advantage: The Birth of the VCR Industry," *California Management Review*, Summer, 1987, 29 55-76.
- Steinmueller, W. Edward, "Microeconomics and Microelectronics: Economic Studies of Integrated Circuit Technology," unpublished doctoral dissertation, Stanford University, 1987.
- , "Industry Structure and Government Policies in the U.S. and Japanese Integrated Circuit Industries," in John B. Shoven, ed., *Government Policies Toward Industry in the United States and Japan*, Cambridge University Press, forthcoming.
- , "International Joint Ventures in the Integrated Circuit Industry," in David C. Mowery, ed., *International Collaborative Ventures in U.S. Manufacturing*, American Enterprise Institute, Ballinger forthcoming, 1988.
- National Science Foundation, National Science Board, *Science Indicators 1985*, Washington, 1985.

## Gender Difference: The Role of Endogenous Preferences and Collective Action

By ELAINE MCCRATE\*

For the last twenty-five years, our society has scrutinized relationships between women and men to an unprecedented degree. In this discussion—at once a positive description of behavior and a normative evaluation of family life—neoclassical economics has contributed important insights concerning the connection between the labor market and the family. However, the neoclassical inquiry has proceeded in remarkable isolation from interdisciplinary and popular debates which have focused on the issue of power relations between men and women—a stance which ultimately restricts the explanatory scope of the analysis.

The invisibility of power is particularly characteristic of the “trade” metaphor for marriage (Gary Becker, 1981; myself, 1987; Nancy Folbre, 1986). The exclusive focus on trade reduces social power to mere purchasing power, which is exercised over technical resources or consumption goods, not people. The transaction cost analysis of the family (Robert Pollak, 1985, and Paula England and George Farkas, 1986) has a richer conception of power relations between family members who are locked in bilateral monopolies, but here power seems to be little more than a phenomenon endemic to long-term personal relationships.

<sup>†</sup>*Discussants:* Gary S. Becker, University of Chicago; Paula England, University of Texas-Dallas.

\*Assistant Professor of Economics, University of Vermont, currently at UCLA, Center for Afro-American Studies, Los Angeles, CA 90024. I thank Samuel Bowles, Nancy Folbre, Manuel Pastor, and participants in the Greater Los Angeles Political Economy Seminar for helpful discussions.

In this paper I advance two different propositions concerning the mechanisms and consequences of power between men and women, and use them to consider some concrete questions about the household. First, the mechanisms of power include collective action (a point argued in other contexts by Douglass North, 1981; Amartya Sen, 1977; Albert Hirschman, 1985; Samuel Bowles and Herbert Gintis, 1986; and Heidi Hartmann, 1976). Second, the consequences of power include the social construction of gender. A discriminatory economic system produces men and women whose socially differentiated capacities for performing and enjoying various types of work are much more distinct than their innate endowments. In other words, preferences and productive abilities are endogenous. (In this paper, I borrow from the work of Hirschman, Sen, Gintis, 1972, and many feminist authors such as Alison Jaggar, 1983, to explore the endogeneity of preferences.)

I use two questions to motivate consideration of these claims. The first concerns the informal marriage contract: why has men's participation in household labor and child-care been so slow to change as women's market income has increased? Time-budget studies generally show that men's household labor is not very responsive to women's market labor. At most, men took on about two more hours of household responsibilities per week in the 1970's when women's wage labor increased dramatically (Ellen Fried and Susan Settergren, 1986; C. Russell Hill and Frank Stafford, 1980; Hartmann, 1981). The second question concerns the formal marriage contract: why are there so many re-

strictions on marital freedom of contract? These include the requirement of a lifelong partnership, the lack of options in the contract's provisions (including those concerned with possible dissolution), and the prohibition of homosexual marriages. This last feature of marriage contracts is so fundamental that economists have apparently been satisfied to assume it.

I consider both of these questions in turn, arguing that they pose special problems for the two neoclassical approaches, and then explain how a consideration of endogenous preferences and collective action can better illuminate these issues.

### I. The Domestic Sexual Division of Labor

The trade analysis of marriage offers two contradictory predictions concerning trends in the domestic sexual division of labor when women's market incomes increase. The first prediction results from the extension of the Coase Theorem to the question of divorce (H. Elizabeth Peters, 1986). If women's nonmarital wealth increases but total marital wealth still exceeds the sum of the couple's wealth as single persons, and if spouses negotiate over the allocation of wealth (broadly interpreted to include utility from different types of work), then one would expect a reallocation of household labor among couples who remained married. For this hypothesis, the rigidity in the domestic sexual division of labor is an anomaly.

The alternative prediction arising from the trade analysis follows from Becker's (1981, 1985) elaborations on the theory of comparative advantage in marriage. A sex-typed pattern of human capital investment and specialization in work is based on women's comparative advantage in childbearing and rearing, given constant or increasing returns to scale in production with specific human capital (household vs. nonhousehold). If women have just a slight advantage in the production of children, then men will specialize completely in market work and market investment. An important implication is that small differences in comparative advantage will produce big differences in specialization,

so even large changes in women's nonmarital incomes might not be enough to induce greater household labor among men.

The question remains as to the source of comparative advantage by gender: while sex discrimination could account for it, Becker (1985) strongly suggests that even on some counter-earth with no discrimination in market work, there would still be a pronounced sexual division of labor for biological reasons. This, of course, involves some very highly disputed notions of biological determinism. It also assumes that there are no significant advantages to children regularly interacting with two adults rather than one. Thus Becker's arguments about complete specialization can in principle explain the rigidity in the domestic division of labor, but they depend largely on some unverified assumptions.

Transaction cost theory likewise has difficulty explaining the inertia in men's household roles. Because both spouses invest heavily in relationship-specific capital and make long-term contractual commitments, marriage is characterized by bilateral monopoly. Accordingly, the concept of bargaining power as understood in standard game theory is central to the analysis. While this expands the notion of power beyond that adopted in the conventional trade approach, it presents new problems, which are evident in the question of housework. In most games, when threat points move, the solution changes. One would therefore expect that wives would negotiate greater household labor by men as women's nonmarital opportunities improve. Yet this appears not to have happened. England and Farkas' important observation that women traditionally invested more in the highly idiosyncratic emotional component of marriage capital compounds the problem: if women's market work has reduced the time they invest in "emotional capital," then one would expect even greater renegotiation of men's domestic roles.

The problem common to both standard game theory and trade theory is the assumption that all the relevant information for decision making is included in relative prices,

productivities, or the location of threat points. Both approaches fail to consider the social identities of the actors, which may impede adjustment to changed economic circumstances. The symmetry property of most games, which requires that the players be able to change their objective positions without altering the outcome of the game, does not hold in marriage.

Recognition of the endogeneity of preferences can help resolve the problem about housework. This view begins by positing 1) the social production of people, that is, of their technical and appreciative capacities, and 2) the ability of people to reflect upon and change their own preferences. The first assertion distinguishes this view from liberal political theory and neoclassical economic theory, which regard individuals as discrete bundles of preferences and endowments constituted prior to social life. Instead, I maintain that individuals are produced by social activity, which includes their own actions in the context of their social experiences (Bowles-Gintis).

The second premise for the endogeneity of preferences is Sen's and Hirschman's observation that people have tastes not just about external objects or other people, but also about themselves: in other words, about their identities. Identity is what these authors have called a "metapreference" or "value." While we may not quibble about tastes over many of the choices we make, we do struggle regularly with ourselves over who we are and who we want to be: we have second-order preferences about our preferences concerning such fundamental issues as manhood or womanhood.

A male-dominant society limits opportunities for women and men to explore and develop their identities. Even independently of manipulation by media and family, women's preferences will develop differently from men's. First, if women or men are so unfamiliar with some of the other sex's work as to be fundamentally uncertain about what it is, they cannot rank it in a preference ordering. Second, if women have reasonable knowledge of all activities, they will learn through observation of others, and through

their own experience, that the most promising route to success or fulfillment involves investing in feminine and maternal identity. Men learn correspondingly about masculinity. Women therefore choose to *learn to prefer* mothering over auto mechanics for the same reason that one would choose to learn to enjoy winter rather than summer sports in cold climates: the expected payoff is higher. This differs from Becker and George Stigler's (1977) consideration of investment in "consumption capital" by recognizing that the investment is embodied in people, is not capable of disinvestment, and thus fundamentally changes one's identity.

Hence men and women rationally make large and long-term investments in sex-typed preferences or identities, developing very different capacities for tastes, and severely restricting the possibility that they may elect at some future point to invest in significantly different tastes. Men's slow entry into household labor is now more comprehensible: given such costly, long-term investments in masculine identity, it would be surprising indeed if men began to do more child care just because of a change in relative prices or productivities, or because of movement of the threat points. A redivision of family labor would fundamentally threaten the value of costly investments in gender identity.

## II. The Legal Marriage Contract

Conventional trade analysis has said very little about the peculiarities of the marriage contract. Possibly the objective of the contract is the efficient production of children; however, one must then wonder about the failure to require or enforce child support from financially able divorced fathers (Barbara Bergmann, 1986). Becker (1981) briefly suggests that the contract protects women whose investments in their families have reduced their ability to prosper outside the family. However, it is peculiar that the traditional marriage contract would include a no-divorce provision to protect women, when the costs of divorce have in part been created by the contract itself. The marriage contract reduced women's ability

to maintain a comparable standard of living outside marriage by restricting their rights to their own property and earnings, their access to credit, their mobility, etc., and by not recognizing the value of their contribution to their husbands' human capital (Lenore Weitzman, 1981).

Transaction cost analysis offers a much more detailed account of the marriage contract, which is similar to the Hobbesian theory of the state. Because the outcome of bilateral bargaining is unpredictable, and because not all conflicts can be foreseen, both spouses assent to external imposition of a long-term contract to facilitate investment in relationship-specific capital. The state and church act to reduce transaction costs, that is, to avoid the personal and social chaos arising from unresolved marital conflict.

However, the transaction cost account cannot explain why the marriage contract has so overwhelmingly resolved conflict in men's favor. Why under the principle of coverture was women's identity subsumed under men's? Why has dissolution always been structured so as to increase men's consumption but lower women's? Furthermore, why is marriage prohibited among homosexuals, since there is no identifiable transaction cost associated with homosexuality? Many of these rules do not seem to be cost minimizing, but rather cost creating, insofar as resources must be allocated to their advocacy and enforcement. Indeed, the marriage contract appears to be a set of rules advanced by a powerful interest group.

Neoclassical theory has demonstrated the conditions that must hold for such effective collective action. First, there must be a group or coalition which controls enough resources to promote successfully its own interests. Second, the group must have means to overcome the free-rider problem associated with collective action. Third, it must have means to maintain cohesion when private incentives induce individuals to renege on collective agreements.

The satisfaction of the first condition, men's substantially greater control of resources, is obvious. The earlier discussion of people's participation in creating their own

identities, given the social circumstances they experience, elucidates the process by which the other two conditions might be met. Men and women can overcome the free-rider problem because collective action reinforces their gender identity. As Hirschman has noted, group identity is available predominantly by acting as part of a group. Thus rational individuals are often quite activist about their own gender roles, and about the roles of others, over issues ranging from proper attire to proper work for men and women. Men may be more effective than women at developing coalitions over these concerns: they certainly have more opportunities to learn collective behavior than do women, ranging from sports events to legislative sessions.

Several means exist to enforce collective norms. Culture is one of them: men and women who deviate from the norms (for example, men who perform housework) are punished. Law is another. Law may function as a form of precommitment to maintain group solidarity when individuals have private incentives to violate group norms. (For example, laws restricting women's participation in wage labor may offset husbands' incentive to acquire more money income through wives.) Additionally, law shapes values.

Consequently, people can act in coalition to create laws which have more consequences than simple rule enforcement. They will attempt as groups to secure legislation in their economic self-interest, or to reinforce their identities.

These aspects of marital law are most relevant for explaining the unusual provisions of the marriage contract, including the prohibition of homosexual marriage. Homosexuality threatens gender identity. It threatens the assumed necessity of a sexual division of labor which pairs women with men under implicit and explicit terms which are often inimical to their economic self-interest.

In conclusion, this analysis has surely raised as many questions as it has answered. However, the questions are fundamental, and have been greatly neglected by economists



Much more philosophical, empirical, and historical work will be necessary to assess the relevance of all three approaches in answering them.

## REFERENCES

- Becker, Gary S., "Human Capital, Effort, and the Sexual Division of Labor," *Journal of Labor Economics*, January 1985, 3, S33-S58.
- , *A Treatise on the Family*, Cambridge: Harvard University Press, 1981.
- and Stigler, George J., "De Gustibus Non Est Disputandum," *American Economic Review*, March 1977, 67, 76-90.
- Bergmann, Barbara R., *The Economic Emergence of Women*, New York: Basic Books, 1986.
- Bowles, Samuel and Gintis, Herbert, *Democracy and Capitalism*, New York: Basic Books, 1986.
- England, Paula, and Farkas, George, *Households, Employment and Gender: A Social, Economic and Demographic View*, New York: Aldine, 1986.
- Folbre, Nancy, "Hearts and Spades: Paradigms of Household Economics," *World Development*, February 1986, 2, 245-55.
- Fried, Ellen Shapiro and Settergren, Susan, "The Effects of Children on Wives and Husbands' Allocation of Time," mimeo., Research Triangle Institute, 1986.
- Gintis, Herbert, "Consumer Behavior and the Concept of Sovereignty: Explanations of Social Decay," *American Economic Review Proceedings*, May 1972, 62, 267-78.
- Hartmann, Heidi, "The Family as Locus of Gender Struggle: The Example of Housework," *Signs*, Spring 1981, 6, 366-94.
- , "Capitalism, Patriarchy, and Job Segregation by Sex," in Martha Blaxall and Barbara Reagan, eds., *Women and the Workplace: The Implications of Occupational Segregation*, Chicago: University of Chicago Press, 1976, 137-70.
- Hill, C. Russell and Stafford, Frank P., "Parental Care of Children: Time Diary Estimates of Quantity, Predictability, and Variety," *Journal of Human Resources*, Spring 1980, 15, 219-39.
- Hirschman, Albert O., "Against Parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse," *Economics and Philosophy*, January 1985, 1, 7-21.
- Jaggar, Alison M., *Feminist Politics and Human Nature*, Sussex: Rowman and Allanheld, 1983.
- McCrate, Elaine, "Trade, Merger and Employment: Economic Metaphors for Marriage," *Review of Radical Political Economics*, Spring 1987, 19, 73-89.
- North, Douglass, *Structure and Change in Economic History*, New York: Norton, 1981.
- Peters, H. Elizabeth, "Marriage and Divorce: Informational Constraints and Private Contracting," *American Economic Review*, June 1986, 76, 437-54.
- Pollak, Robert A., "A Transaction Cost Approach to Families and Households," *Journal of Economic Literature*, June 1985, 23, 581-608.
- Sen, Amartya K., "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs*, Summer 1977, 6, 317-44.
- Weitzman, Lenore, *The Marriage Contract: Spouses, Lovers and the Law*, New York: Free Press, 1981.

# Tied Transfers and Paternalistic Preferences

By ROBERT A. POLLAK\*

Why do parents make inter vivos transfers to their children and leave them postmortem bequests?<sup>1</sup> Gary Becker's notions of "altruism" (1981, ch. 8)—by which he means that children's utilities are arguments of their parents' utility function—provides one explanation. Denoting the children's utility functions by  $U^i(c_i)$ , where  $c_i$  denotes consumption by child  $i$ , the preferences of parents with two children can be represented by a utility function of the form  $W[c_p, U^1(c_1), U^2(c_2)]$  where  $c_p$  is the parents' own consumption.<sup>2</sup> In the altruistic model, parents' sole motive for intergenerational transfers is to increase their children's utility. In other models, however, parents may have nonaltruistic as well as altruistic motives for transferring resources to their children.

The literature on economic development emphasizes old-age support as a motive for fertility and, to a lesser degree, as a motive for providing children with human capital as part of an explicit or implicit intergenerational contract. When human capital formation is the primary focus of the analysis, as in discussions of education and earnings, it is useful to decompose inter vivos transfers into "human capital formation" and "other inter vivos transfers." Such a decomposition can mislead in discussing intergenerational transfers, however, because it obscures the fact that the provision of human capital by parents constitutes an intergenerational transfer.

Laurence Kotlikoff and Avia Spivak (1981) analyze another old-age support model, one in which the family operates as an "incomplete annuities market." In their model, children make regular transfers to their aging parents, and the share that each child contributes to the parents determines his or her share of the parental estate. Although the prospect of old-age support may be an important motive for "downstream" intergenerational transfers (i.e., from parents to children) in some societies, in the United States today "upstream" transfers appear too small and too uncertain to make this motive credible.

To explain downstream transfers in the United States today, economists have investigated other models in which parents have selfish as well as selfless motives. For example, B. Douglas Bernheim, Andrei Shleifer, and Lawrence Summers propose a model in which parents use the prospect of bequests to exact services from their children: "we envision a testator who, though altruistic, is also affected by actions taken individually by a number of potential beneficiaries (he may, e.g., enjoy receiving attention from his children)" (1985, p. 1046). Bernheim et al. assume that such actions increase parents' utility and decrease children's utilities. In their model the children's utility functions become  $U^i(a_i, c_i)$ , where  $a_i$  denotes services the  $i$ th child provides the parents, and the parents' utility function becomes  $W[c_p, a_1, a_2, U^1(a_1, c_1), U^2(a_2, c_2)]$ . To measure these services, Bernheim et al. use frequency of contact (i.e., visits plus telephone calls) between parents and children.

This approach expands the concept of child services to include those provided by adult children who live outside the parents' household in an attempt to explain bequests and inter vivos transfers. "Child services" originally appeared in discussions of fertility and the allocation of resources to young children living with their parents, and thus tends to evoke the joys of young parenthood

\*University of Pennsylvania, Philadelphia, PA 19104, and University of Washington. I am grateful to the National Science Foundation and the National Institutes of Health for financial support, to Gary Becker, Samuel Preston, David Stapleton, and Paul Taubman for helpful comments, and to Judith Farnbach for editorial assistance.

<sup>1</sup>Even if bequests are unplanned, as some versions of the life cycle savings model assume, inter vivos transfers must be intentional.

<sup>2</sup>I ignore the possible dependence of the children's utility on their own children's utility, etc., because it is not relevant to the issues discussed in this paper.

(the pitter-patter of little feet). In that context there are two versions of the child services or "child quality" model. The first focuses on a number of narrowly defined child-related commodities produced by the household's technology (for example, musical skills, mathematics grades, the negative of the number of ear aches). The second focuses on a single broadly defined "commodity" that is best interpreted as the value of a separable component of the parental utility function defined over the vector of narrowly defined child commodities. The boundary separating the young children to whom the child quality model applies from the adult children to whom the altruism model or the adult version of the child services model applies is a broad band rather than a narrow line; college students, for example, are an ambiguous case. As in the story of the six blind men each describing a different part of the elephant, one might believe that the child quality model and the other models each capture an important part of the situation without believing that any one of them captures it all.

In this paper I propose the "paternalistic preferences" model, which accounts for aspects of downstream transfers, whether *inter vivos* or bequests, that neither Becker's pure altruism model nor the child services model can explain. The most crucial of these aspects is the prevalence of conditional or "tied" transfers. In the paternalistic preferences model, parents' utility depends directly on their children's consumption patterns as well as on their children's utilities, so that the parents' utility function is of the form  $W[c_p, c_1, c_2, U^1(c_1), U^2(c_2)]$ , where the  $c$ 's must now be interpreted as consumption vectors.<sup>3</sup> The fundamental insight of the paternalistic preferences model is that parents care about their children's consumption patterns even after the children are grown and have left home. The model thus occupies the middle ground between the child quality models and the other models.

Intergenerational transfers are often tied to the child's consumption of particular goods and services. Three examples illustrate the point. (i) Many parents spend substantial resources on their children's college educations. Suppose your college-age daughter announces that she would rather use the money for something else, for example, to contribute to Green Peace or to buy a Mercedes. As a parent I would find such an announcement distressing, and I suspect most parents would; my guess is that in most families the daughter would not get the money for Green Peace or for the Mercedes. Why? (ii) Anecdotal evidence suggests the importance of financial help from parents in providing down payments for home purchases; some confirmation—it has been suggested that "data" is the plural of anecdote—is provided by statistics from the National Association of Realtors (1986, p. 25) showing that gifts are the source of 19 percent of the funds used for down payments on first houses. Suppose you were planning to give your son the money for the down payment on his first house, but that he would rather use the money for something else, for example, to contribute to Green Peace or to buy a Mercedes. I suspect that few of the parents who were willing to give their son the money for the down payment would give him the money for Green Peace or for the Mercedes. Why? (iii) Anecdotal evidence suggests that parents with substantial wealth are sometimes concerned that their children will misuse, squander, or waste resources that are transferred to them. To mitigate these concerns, parents sometimes establish trust funds that limit children's control over resources and thus their opportunities to misuse them. What lawyers call "spendthrift trusts" are specifically designed to restrict children's control over resources, often well beyond the children's minorities and the parents' lifetimes. Why?

The altruistic model and the child services model find tied transfers anomalous.<sup>4</sup> In the

<sup>3</sup>Formally, paternalistic preferences generalize the child services model by including actions that have positive as well as negative effects on the children's utilities. For expositional ease, I focus on actions with positive effects.

<sup>4</sup>This is overstated. At least for education, tied transfers are compatible with a multiperiod version of the altruistic model in which parents implicitly insure their children against income shortfalls and, hence, have an interest in their children's self-sufficiency. Insurance and

altruistic model, parents want each child to choose the consumption pattern that maximizes his or her utility, subject to the appropriate resource constraints. In the child services model, parents want to exact services from their children, but once the vector of services provided to the parents is fixed, parents want each child to choose the consumption pattern that maximizes his or her utility, subject to the appropriate resource constraints. Thus, these models cannot explain why parents are willing to transfer resources to their children only when those resources are committed to particular uses. In the paternalistic model, parents want the child to choose a consumption pattern that reflects not only the child's utility, but also the parents' desire that the child consume certain goods and eschew others. Thus, the paternalistic model explains tied transfers.

What motives might underlie paternalistic preferences? First, parents might want their children to attend college or own a house because it gives the parents pleasure or satisfaction, independent of their children's preferences. This might reflect the parents' own values and aspirations, or their concern with status ("my daughter the doctor"). Second, parents might believe that attending college or owning a house is in the child's true, long-run interest ("when you're older, you'll thank me for this"). In this case the parents' motives are "altruistic" in the ordinary language sense (i.e., concerned with the welfare of others), although not in the Beckerian sense of respecting their children's preferences.

Parents attempt to influence their children's consumption patterns in two ways: by altering the preferences of children and by altering the opportunities of children. Although economists rely heavily on the assumption that preferences are fixed and immutable, many would concede that preferences can be influenced by example and

persuasion, and that families play a crucial role in preference formation. Parents may alter opportunities in two ways. In the household production framework, the technology by which market goods are transformed into commodities is a determinant of a child's opportunities. Any analysis based on the acquisition of tastes can be reformulated in terms of the acquisition of technology. Second, parents may alter opportunities by subsidizing the consumption of certain goods—that is, by offering tied transfers.

The phrase "paternalistic preferences" is familiar from public economics, where it appears in the analysis of "merit wants" and "merit goods." Parallels between the family and society are commonly drawn in both economics and political theory. Paul Samuelson's classic paper on social indifference curves (1956), which used the Bergson-Samuelson social welfare function in the context of intrafamily allocation, was the first to identify a precise analytic correspondence. In the light of these parallels, it is not surprising that the reason we, as a society, provide the poor with access to certain goods (for example, through food stamps and medicaid) rather than with the resources to purchase goods corresponds to the reason we, as parents, provide our children with tied transfers. To the extent that social policy reflects concern for the children of the poor rather than for the adult poor, the two meanings of "paternalistic" intertwine and reinforce each other. Nevertheless, the word paternalistic has unfortunate connotations because of its not entirely coincidental gender coding.

To construct an intergenerational allocation model, we must combine assumptions about preferences with assumptions about constraints (which I shall not discuss) and assumptions about equilibrium or solution concepts (to which I now turn). In my paper (1985), I argue that Becker completes his altruistic model by implicitly assuming that the altruist (the husband-father-patriarch-dictator) has the power to impose his preferred allocation on the other family members, subject only to the constraint that none of them can be made worse off than they

---

inalienability do not, however, provide a satisfactory account for houses and other tied transfers. Later in the paper, I discuss anecdotal evidence suggesting that these motives do not provide a complete account of the transfers associated with education.

would be if they withdrew from the family. This solution depends crucially on the implicit assumption that it is the altruist rather than the rotten kid who, by the rules of the game, can face the other players with a take-it-or-leave-it choice. If the altruist can make the final offer, then he gets the entire utility surplus; if the rotten kid can make the final offer, he grabs it all. Donald Cox (1987) uses this approach to analyze inter vivos transfers in a child services model, treating the problem as one in which parents maximize utility subject to the constraint that no child can be made worse off than he or she would be by withdrawing from the family.

Bernheim et al. formulate the intergenerational allocation problem as a noncooperative game involving parents and two or more children. They assume parents want to leave their estate to their children (so that leaving everything to other relatives, to friends, or to charity is not a credible threat), but that parents are indifferent among alternative divisions of their estate among their children (so that leaving everything to one child and disinheriting the others is a credible threat). Bernheim et al. also assume the children compete with one another for parental favor (so the children do not form a coalition). Under these assumptions, they show that the solution is one in which the parents get the entire utility surplus. The principal contribution of Bernheim et al. is their formulation of the intergenerational allocation problem as a noncooperative game rather than the particular formulation they propose. Indeed, the appeal of the particular Bernheim et al.-Bertrand formulation will be greatest to those who expect duopoly to yield competitive prices. The analogy with game-theoretic models of duopoly suggests two lessons: the vastness of the range of alternative modeling assumptions and the sensitivity of solutions to these assumptions. Thus, the Bernheim et al. analysis is unlikely to be the last word on the subject.

The paternalistic preferences model presents similar modeling choices. The simplest formulation is one in which parents maximize their own utility, subject to the constraint that no child can be made worse off than he or she would be outside the family.

More complex formulations would use cooperative or noncooperative game theory.

How do the implications for investment in human capital differ between the altruistic model and the parental preferences model? The maximizing version of the altruistic model yields sharp results, at least in special cases. Suppose children value education only for its economic returns (so that they derive neither utility nor disutility from attending school or from the prospect of being an educated man or woman). Assuming positive inter vivos transfers or bequests to all children, Becker and Nigel Tomes (1976) show that the altruistic model implies a two-stage allocation process. First, parents invest in each child's schooling until the marginal return equals the return on financial assets; then they make transfers and bequests among their children to obtain the desired income distribution.<sup>5</sup>

When parents value their children's educational attainments, the maximizing version of the paternalistic preferences model implies that under the assumptions of the previous paragraph (i.e., children value education only for its economic returns; parents make positive transfers or bequests to all children), the marginal return on education should be less than the return on financial assets. Some anecdotal evidence supports this conclusion, although there is little systematic evidence. The anecdotal evidence I have in mind involves the education of children from upper-income families. Instead of focusing on years of schooling, consider expenditure on schooling. Private colleges are expensive: tuition at the University of Pennsylvania, an Ivy League school, is \$11,976; at Pennsylvania State University, a public institution, tuition for an in-state student is \$3,292. If children from upper-income families are more likely to attend private colleges and

<sup>5</sup>Hence, if parents place any value on equality in the distribution of income among their children, they will compensate children with lower earnings by providing them with greater transfers and/or bequests. Furthermore, assuming parental preferences exhibit "equal concern" for all their children, the parents will use transfers and bequests to offset fully differences in earnings.

universities than children of equal ability from less wealthy families, then we must conclude that either (a) wealthier families "overinvest" in education (i.e., invest beyond the point at which the marginal returns are equal to the return on financial assets), or (b) less wealthy families "underinvest" in education.<sup>6</sup> Private secondary school education poses the same issue: if children from wealthy families are more likely to attend such schools than children of equal ability from less wealthy families, this indicates either that the returns to such education are greater than the returns on financial assets or that some of the benefit from such education accrues to the parents rather than to the children—in short, that such education can only be understood in terms of paternalistic preferences.<sup>7</sup>

Tied transfers are important, yet the desire of parents to prevent their children from squandering, misusing, or wasting resources

is unintelligible in the altruistic model in which parents passively accept their children's preferences. It is also unintelligible in the child services model, in which parents accept their children's preferences but also want their children to provide services. It is fully comprehensible in a model in which parents evaluate their children's consumption patterns according to their own paternalistic preferences.

## REFERENCES

- Becker, Gary S., *A Treatise on the Family* Cambridge: Harvard University Press 1981.
- and Tomes, Nigel, "Child Endowments and the Quantity and Quality of Children," *Journal of Political Economy* August 1976, 84, S143–62.
- Bernheim, B. Douglas, Shleifer, Andrei and Summers, Lawrence H., "The Strategic Bequest Motive," *Journal of Political Economy*, December 1985, 93, 1045–76.
- Cox, Donald, "Motives for Private Income Transfers," *Journal of Political Economy* June 1987, 95, 508–46.
- Kotlikoff, Laurence J., and Spivak, Avia, "The Family as an Incomplete Annuities Market," *Journal of Political Economy*, April 1981, 89, 372–91.
- Mauss, Marcel, *The Gift* (1925), New York: W. W. Norton, 1967.
- Pollak, Robert A., "A Transaction Cost Approach to Families and Households," *Journal of Economic Literature*, June 1985, 23, 581–608.
- Samuelson, Paul A., "Social Indifference Curves," *Quarterly Journal of Economics* February 1956, 70, 1–22.
- National Association of Realtors, *The Homebuying and Selling Process*, Washington, 1986.

<sup>6</sup>By equal ability I mean identical "earnings production functions"—i.e., the same equation determines each child's earnings as a function of his or her education.

<sup>7</sup>Throughout this paper I have avoided the term "gifts," which sociologists and anthropologists use, in favor of the more neutral "transfers." The basic work on gifts and "gift economies" (Marcel Mauss, 1925), emphasizes that "gift exchanges" (a telling phrase) are governed by rules and obligations: to make the gift, to receive the gift, and to reciprocate. Even in the gift economy, there is no free lunch. The relationship between practices in "primitive" gift economies (for example, the potlatch of the Kwakiutl and other Northwest Coast Indians and the Kula ring of the Trobriand Islanders) and intrafamily transfers in advanced industrial societies is unclear. Whether donors and recipients characterize a particular transfer as a gift may signal whether they believe it carries with it reciprocal obligations. For example, if a parent provides \$10,000 for a child's down payment on a house, it is called a gift. If a parent provides \$70,000 to send a child to college, it is not called a gift.

# Risk, Private Information, and the Family

By MARK R. ROSENZWEIG\*

In the last fifteen years, our understanding of the behavior of households in both high- and low-income countries has increased significantly. Econometric studies contributing to this body of knowledge, concerned with the determinants of such behavior as labor supply, fertility, health and food consumption, however, have generally taken the structure of the household as exogenously given. While some progress has been made concerning how changes in the legal structure alter patterns of family formation and breakup, the wide variety of family organizations observed across countries of the world or the evolution of family structure within countries cannot be readily explained by existing models. While some economists have suggested that organizations based on kinship can be understood in terms of transaction economies (Yoram Ben-Porath, 1980; Robert Pollak, 1985), there are now few precise implications of this approach, and little or no evidence of its usefulness for predicting or explaining the existence of any particular family structure in specific settings characterized by their natural endowments and/or legal structure.

In this paper I discuss some recent studies of family and household organization in one specific context, rural India, that have sought to formulate a model of household structure based on the need for individuals in low-income, private information settings to protect themselves against intertemporal fluctuations in resources arising from the natural vagaries of water supply. I also present some new evidence on the relationship between family arrangements and measured risk characteristics of the agricultural environment in these settings. While many economists have highlighted risk and information consideration in the study of such formal rural institutions as sharecropping, "perma-

nent" servitude, and contractual interlinking, studies of these individual contractual arrangements have ignored the family itself as a risk-mitigating institution. Moreover, much of this literature has been solely concerned with *ex ante* measures to reduce risk—contractual instruments or production factors that reduce the effects of variability in the exogenous production input rainfall on income. Mechanisms serving to preserve consumption stability *ex post* in the face of variability in realized income have been neglected.

## I. *Ex Ante* and *Ex Post* Arrangements to Mitigate Risk in Stationary Agricultural Environments

There are three important characteristics of agricultural production that may influence the form of organizations in agricultural environments. First, one important production factor, land, is immobile and heterogeneous. Second, another important production input, rainfall, varies substantially across production cycles and across space. Third, rainfall is spatially covariant. As a consequence, agricultural incomes vary over time, are imperfectly correlated across production units, but are more correlated among production units the more proximate are such units. If the income  $\pi_t^{ij}$  for a production unit  $i$  in an environment  $j$  defined by the commonality of its rainfall state in each period  $t$ ,  $\varepsilon_t^j$ , is given by

$$(1) \quad \pi_t^{ij} = \psi(\varepsilon_t^j, A_t^{ij}),$$

where  $A_t^{ij}$  is a vector of the  $i$ th household's fixed (within the production period) production inputs and  $\psi_{\varepsilon, A} \neq 0$ , then because of the heterogeneity of  $A_t^{ij}$ ,  $\text{cov}(\pi_t^{ij}, \pi_t^{kj}) \neq 0$ . The incomes of production units within a common weather environment may not covary perfectly because of heterogeneity in production factors; the incomes of households in

\*University of Minnesota, Minneapolis, MN 55445.

different environments also covary, but less strongly because of both factor heterogeneity and the imperfect spatial covariance of states of nature (i.e.,  $\text{cov}(\pi_t^{ij}, \pi_t^{il}) < \text{cov}(\pi_t^{ij}, \pi_t^{kj})$ ). All attempts by households to reduce the impact of  $\epsilon_t$  on realized incomes  $\pi_t$  in (1) are *ex ante* risk measures.

The variability in rainfall implies that risk-averse agents will devote resources to stabilizing their consumption. Less than perfect correlations in agricultural incomes implies that individuals can reduce the volatility of their consumption *ex post* by pooling incomes across space. The (positive) spatial covariance of weather, however, implies that the gains from pooling rise as the geographical scope of risk sharing increases. On the other hand, costs of monitoring and enforcement of such *ex post* insurance schemes, necessitated by the prospects of moral hazard, also may rise with distance. The tradeoff between diversification gains and contractual costs associated with the distance of such insurance-like transfers may be reduced, however, if such arrangements are contracted among family members, individuals who both have specific knowledge about each other and who may also "care" about each other's welfare. Robert Lucas and Oded Stark (1985), for example, present some evidence from Botswana that the remittances of household temporary migrants (males) exhibit risk-reducing properties.

In rural India, however, the migration of males, temporary or permanent, is a relatively small phenomenon. Farm households appear to be extended vertically, not spatially, at least with respect to males. That is, fathers, mothers, and adult sons (and their immediate families) jointly reside in a single household and jointly cultivate the family land. My article with Kenneth Wolpin (1985) showed that consideration of *ex ante* risk reduction could account for this form of household structure. We argued that land heterogeneity and the stationarity of weather in a regime of unchanging technology would create payoffs to farming experience specific to plots of land in terms of the ability to mitigate profit losses associated with bad weather. Given the specificity of experiential returns, we showed that fathers and children would have incentives to farm together on

the family land, with the children benefiting from the experience of the elder and gaining experience on the family land. The (male) children would inherit the family land at the death of the head, since the sale of such land by the elder to any other anonymous agent would result in a capital loss (the value of experience) to the *family*. The model thus simultaneously accounts for the scarcity of land sales, the immobility of landholders, and the particular intergenerational structure of the households in settings characterized by land heterogeneity, variable rainfall and stationarity.

Wolpin and I obtained evidence from a national longitudinal probability sample of Indian farm households that more-experienced farmers suffered less in terms of farm profitability under adverse weather conditions by estimating a variant of (1), although we could only employ age as a proxy for specific experience. We also showed that sales of land were significantly less likely when the family was intergenerationally extended, for given weather conditions, mean farm profitability and the schooling attainment of the head.

The Rosenzweig-Wolpin model ignores the gains from spatial income diversification and, in particular, the role of daughters and daughters-in-law in contributing, indirectly, to *ex post* income smoothing. Marriage is in part a means by which households not necessarily proximate to each other create or solidify family ties that traverse space, and, indeed, almost all rural mobility in India is accounted for by women who move for the purpose of marriage. In my forthcoming paper, I obtained evidence based on longitudinal data from rural South India, described below, that the number of daughters-in-law in a household increases the effectiveness of cross-household transfers (net of dowry and any marital gifts) in mitigating the effects of variability in household income, as do to a lesser degree household migrants and the nonresident siblings of the household head. My paper with Stark (1987) also showed that the distance of the origin households of resident marital partners contributed directly to reducing the intertemporal variance in household food consumption for given income variance. Gross transfers, over 60 per-



cent of which (in value terms) originated outside the villages in which the sampled households resided, accounted for about 10 percent of agricultural profits (net of the value of family labor) on average. While the use of credit played a greater role in *ex post* risk mitigation than did transfers, the availability of credit was highly dependent on the performance of the local economy, while the volume of transfers did not depend significantly on local incomes. The extra-village ties facilitated by exogamy thus provide better insurance against severe common shocks than does the locally based credit market.

## II. Marital Arrangements and Consumption Smoothing

In my paper with Stark, the role of marital arrangements in contributing to consumption smoothing *ex post* is elaborated and predictions derived and tested concerning assortative mating patterns and the location of marriage, based on risk considerations. We argued that the transfer of family members across spatially separated households facilitates risk sharing, given spatially covariant risks, for two reasons: first, the presence of a member of household *A* in a household *B* provides an incentive for household *A* to be concerned about consumption in household *B* (if *A* cannot control the intrahousehold distribution of resources in *B*), assuming *A* still cares for its former members. Second, if the transferred member from *A* still cares about the members of household *A*, then household *A* has an agent resident in *B* who can monitor *B*, thereby mitigating the effects of moral hazard. With males specializing in farm production and remaining immobile (to exploit specific experiential returns), and females specializing in household production (general skills) and moving to new households, almost every household is monitored by a resident agent and has some degree of *ex ante* and *ex post* risk protection.

Risk considerations provide implications about marital arrangements not readily derived from marriage models concerned solely with income gains. For example, both the standard model of marriage (Gary Becker, 1973) and risk considerations suggest the

optimality of positive assortative mating with respect to permanent characteristics augmenting incomes (wealth), since households engaged in an *ex post* insurance arrangement but unequally matched with regard to permanent wealth would be unequally insured. However, risk considerations imply that households will attempt to match marital partners such that the transitory incomes of the origin households are negatively correlated. This means that the distance between households conveys a positive gain to the extent that distance reduces weather covariances. If wealthy households are less concerned about risk avoidance than are poorer households, then such households may be characterized by more proximate marriages. In contrast, if positive assortative mating by wealth were the only objective, wealthier households might be observed to be engaged on average in longer-distance marriages, given the rarity of high wealth, and thus the necessity of a wider scope of search. Consideration of risk sharing via marriage also suggests that households would tend to build a portfolio of marriages, to diversify risk by arranging marriages in a variety of locations, despite the information-cost advantages of specialization in marital search in one locality.

My paper with Stark, using a special survey of marriages supplementing the Indian village data used in my earlier paper, found evidence consistent with these implications for marriage: within households that had arranged two or more marriages, 94 percent of the marriages did not involve the same origin (daughters-in-law) or destination (daughters) villages; that is, the marriage portfolios were almost completely diversified. Of the 114 marriages for which there was information, moreover, only 8 involved partners from the same survey village. Despite this exogamy, only 14 of the marriages did not also involve partners who were already related by kinship. Marriages thus do not create households but rather solidify pre-existing family ties and neither inheritances nor dowry practices evidently result in the diminution of *family* wealth.

My paper with Stark also obtained evidence on the determinants of marital matches consistent with the risk framework. We found

that farm households with greater amounts of inherited wealth, for given production variability in farm profits, had marriage portfolios with lower mean distances between origin and destination villages, while greater predicted variability in profits, for given wealth, was associated with longer-distance marriages.

### III. Agricultural Income Variability, Specific Experience, and Family Arrangements: Additional Evidence

In this section I present estimates of the interactions between rainfall, farming experience and farm profits, as in (1), to test the basic assumption of the specific experience explanation of intergenerational household extension, and of the effects of exogenous income variability on family structure and the incidence of share tenancy. The data employed are those that were used in my earlier paper and in my paper with Stark based on a ten-year survey of 30 farm households in six villages representing three agroclimatic regions in the semi-arid tropics of India undertaken in 1975 by the International Crops Research Institute for the Semi-Arid Tropics. These areas were untouched by the "green revolution" during the survey period and are characterized by both low levels and erratic distributions of rainfall—the (nine-year) coefficient of variation in profits from crop production for the average farm household in this sample is 139. Control of water for production and avoidance of consumption instability thus represent significant, if not the most significant, problems faced by these households.

Based on a retrospective questionnaire on plot ownership and cultivation administered to all sample households in 1984, the total experience of each household head on all owned plots of land, by dry and irrigated, was computed for each of the ten years of the sample survey. For household  $i$  in year  $t$ , cumulative experience on  $n$  owned plots of dry (irrigated) land is computed as  $\sum \lambda_i^k e_i^{ik}$ , where  $\lambda_i^k$  is the fraction  $k$  of total owned dry (irrigated) land in year  $t$  and  $e_i^{ik}$  is the total years that plot (whether owned or not) had been cultivated by the household head

TABLE 1—EFFECTS OF LAND-SPECIFIC EXPERIENCE, OWNED LAND TYPE AND RAINFALL ON REAL AGRICULTURAL PROFITS, 1975–83

Variable	Estimation Procedure <sup>a</sup>		Mean (Standard Deviation)
	Random Effects	Fixed Effects	
Owned irrigated land (acres)	265 (5.46)	356 (3.26)	2.88 (6.47)
Owned dry land (acres)	38.8 (1.59)	54.0 (1.47)	9.91 (10.5)
Irrigated land $\times$ rain per day (mm) <sup>b</sup>	5.24 (0.85)	5.54 (0.90)	15.2 (40.7)
Dry land $\times$ rain per day	3.73 (1.06)	1.16 (0.32)	48.7 (61.5)
Experience on own dry land (years)	62.1 (2.86)	148 (4.24)	13.0 (11.2)
Experience on own irrigated land (years)	3.14 (0.15)	157 (3.71)	7.73 (11.8)
Experience on own dry land $\times$ rain	-7.45 (2.25)	-6.85 (2.04)	63.0 (59.9)
Experience on own irrigated land $\times$ rain	3.27 (1.02)	4.37 (1.65)	37.3 (61.3)
Age (years)	220 (3.00)	97.8 (0.89)	46.7 (12.0)
Age squared	-2.13 (3.08)	-1.56 (1.50)	2322 (11.96)
Age $\times$ rain per day	5.07 (1.67)	4.48 (1.47)	234 (107)
$F$	32.7	9.11	
$\chi^2$ (error components)	1058	—	
$\chi^2$ (Hausman test)	—	77.8	
d.f. (households) <sup>c</sup>	1165 (148)	1019 (148)	

<sup>a</sup>Asymptotic  $t$ -ratios are shown in parentheses beneath coefficients.

<sup>b</sup>Daily rainfall in the months of July, August and September.

<sup>c</sup>Also included as regressors: mean distance of owned land plots from the household, plot distance interacted with rainfall, the household head's schooling attainment.

up to year  $t$ . The real value of net profits from cultivation in year  $t$  is assumed to be a function of dry and irrigated owned land, total experience on these plots up to year  $t$ , rainfall in year  $t$  (rain per day in three critical months during the rainy season, July, August, and September), and rainfall interacted with both land owned and experience.

Table 1 reports random effects and fixed effects estimates of the profit function. Both sets of estimates indicate that specific experience contributes significantly to agricultural profits, whether on owned dry or irrigated land, and that experience on owned dry land mitigates the influence of rainfall. Moreover, the preferred (based on the Hausman test) fixed effects estimates indicate that the joint effects on profits of age, given specific experience, is not statistically significant. The point estimates indicate that at the sample

TABLE 2—EFFECTS OF PROFIT VARIANCE,  
PROFIT LEVELS AND WEALTH ON  
FAMILY EXTENSION CHARACTERISTICS AND  
INCIDENCE OF SHARE CONTRACTING

Variable/Estimation Procedure	Resident Daughters- in-Law of Head <sup>a</sup> 2SLS <sup>d</sup>	Probability of share Contract <sup>b</sup> Two-Stage ML Probit
Profit variance ( $\times 10^{-8}$ ) <sup>c</sup>	4.94 (1.79)	38.8 <sup>e</sup> (1.60)
Mean profits ( $\times 10^{-3}$ ) <sup>c</sup>	-.053 (1.11)	-.618 <sup>e</sup> (1.49)
Inherited wealth ( $\times 10^{-8}$ )	-9.37 (0.23)	-576 (1.54)
Age of head	.0577 (2.20)	.00589 (0.06)
Age squared ( $\times 10^{-3}$ )	.389 (1.50)	-.0967 (0.11)
Constant	-1.46 (2.28)	-.306 (0.13)
d.f.	180	76
F	5.14	-
$\chi^2$	-	9.96

<sup>a</sup>Sample is from six ICRISAT villages, 1975–83.

<sup>b</sup>Sample is from three ICRISAT villages, 1975–84.

<sup>c</sup>Endogenous variable. Instruments include the means and variances in daily rainfall in July, August, and September interacted with inherited dry and irrigated landholdings.

<sup>d</sup>Asymptotic *t*-ratios in parentheses.

<sup>e</sup>Coefficients are jointly significant ( $F(2,180) = 5.59$ ).

means an increase in rainfall by one millimeter per day increases real agricultural profits by 180 rupees (10 percent). Each additional year of experience on owned dry land increases net profits by 148 rupees at the mean, but reduces the impact of the one millimeter per day rise in rainfall on profits by seven rupees (a 4 percent reduction in the effects of rainfall). Farmers with more cumulative years of experience on dry land thus, *ceteris paribus*, have higher profits and profits less affected by rainfall variability. Experience on irrigated land contributes positively to profits as well but not to a reduction in rainfall-induced profit variability.

Table 1 suggests the existence of gains from continuous cultivation of own land by family members and thus the profitability of cogenational cultivation and family bequests of land. In Table 2, I test whether predicted variability in farm profits and inherited wealth affect the number of co-resident daughters-in-law of the head and the probability that the household engages in one form of *ex ante* risk mitigation, sharecropping. The mean and variance of profits

are computed from the ten years of annual profit data for each farm household. Instruments in the first stage used to predict profit levels and variability include the means and variances in annual daily rainfall interacted with inherited dry and irrigated landholdings. Instruments are used since the actual variability of profits will reflect the household's ability to insure itself *ex post* to the extent that profit variability can be influenced via *ex ante* measures as in (1).

The estimates in Table 2 are consistent with the hypothesis that both household living arrangements and formal contractual choices are influenced by wealth and by income variability. In particular, the number of resident daughters-in-law (= number of married sons), who represent both the extent of intergenerational household extension and of spatial family extension, is greater in households experiencing higher profit variability, for given wealth, while more wealthy households have fewer external ties facilitating *ex post* insurance, given profit variability. The results for sharecropping parallel those for family "extension," with greater variability and lower wealth conducive to a higher incidence of share tenancy and thus a greater degree of *ex ante* protection.

#### IV. Conclusion

In this paper I have discussed how the structure of households and family relationships are shaped by the material conditions of the environment in the context of one specific setting, rural India. Consideration of the income insurance role of family relationships suggests that the growth of formal institutions that serve to mitigate income risk or technological change, which obsolesces experience and makes risk assessment more difficult, may have important effects on the structure of households and on marital arrangements in traditional societies. The general importance of spatially extended family ties in other environments may be obscured, however, by the common survey practice in which the household is the sampling unit; such sampling schemes impede tests of intrafamily and interhousehold risk-sharing hypotheses.

## REFERENCES

- Becker, Gary S., "A Theory of Marriage: Part I," *Journal of Political Economy*, July/August 1973, 81, 813-34.
- Ben-Porath, Yoram, "The F-Connection: Families, Friends and Firms and the Organization of Exchange," *Population and Development Review*, March 1980, 6, 1-30.
- Lucas, Robert E. B. and Stark, Oded, "Motivations to Remit: Evidence from Botswana," *Journal of Political Economy*, October 1985, 93, 901-18.
- Pollak, Robert A., "A Transaction Cost Approach to Families and Households," *Journal of Economic Literature*, June 1985, 23, 581-608.
- Rosenzweig, Mark R., "Risk, Implicit Contracts and the Family in Rural Areas of Low-Income Countries," *Economic Journal*, forthcoming.
- \_\_\_\_\_ and Stark, Oded, "Consumption Smoothing, Migration and Marriage: Evidence from Rural India," University of Minnesota Economic Development Center Bulletin No. 87-11, November 1987.
- \_\_\_\_\_ and Wolpin, Kenneth I., "Specific Experience, Household Structure and Intergenerational Transfers: Farm Family Land and Labor Arrangements in Developing Countries," *Quarterly Journal of Economics*, Suppl. 1985, 100, 961-88.

## HIGH SCHOOL ECONOMICS: IMPLICATIONS FOR COLLEGE INSTRUCTION<sup>†</sup>

### A Report Card on the Economic Literacy of U.S. High School Students

By WILLIAM B. WALSTAD AND JOHN C. SOPER\*

In the 1980's, assessment and critique of American education has taken center stage. A large segment of the public is upset with the educational achievement of precollege students in several content areas. Economics should now be added to the list of failing subjects because the results of our study show a poor performance by many high school students in their knowledge of basic economic concepts.

The study is based on a large, national sample of students who took the second edition of the *Test of Economic Literacy (TEL)* (Soper-Walstad, 1987). The *TEL* is a nationally normed and standardized test of the basic economic understanding of students in eleventh and twelfth grades, consisting of two forms of 46 multiple choice questions. The test questions were based on *A Framework for Teaching the Basic Concepts* (Phillip Saunders et al., 1984). This content guide describes 22 basic economic concepts in four concept clusters—fundamental, microeconomic, macroeconomic, and international—that should be taught in secondary schools to enable students, “by the time they graduate from high school, to understand enough economics to make reasoned judgments about economic questions” (p. 1).

Although economic literacy can be defined and measured in different ways (George Stigler, 1970; W. Lee Hansen, 1977), data

from the norming of the *TEL* provide a comprehensive assessment of the economic literacy of U.S. high school students. The *TEL* was administered as a pre-test to 6,570 students in January 1986. Another 8,205 students took the *TEL* as a post-test in May 1986. Combining the two data sets produced a representative, national sample of 3,031 cases where students had taken the *TEL* as both a pre- and a post-test in one of four courses. This student group will be used for the analysis so that changes in economic literacy across different types of courses can be examined.

Students were classified by type of course based on information from a teacher survey. Of the matched pre- and post-test sample, 50 percent were taking an economics course that used a published high school economics text and focused instruction on basic economic concepts. Students taking courses designated by the teacher as “consumer economics” were 19 percent of the sample. The remaining 31 percent of the students were taking various social studies courses, such as U.S. history or government: 15 percent took social studies courses from teachers who reported including economics in the course; 16 percent took a social studies course without any economics instruction.

#### I. *TEL* Item Performance

The mean percent correct on all the unique *TEL* items by the type of course are reported in Table 1. (For the sake of parsimony, the 46 items on each form were combined and the 15 items that were common to each form were counted only once to produce one 77-item test. The findings from the

<sup>†</sup>*Discussants:* Richard C. Porter, University of Michigan; George H. Borts, Brown University; Donald N. McCloskey, University of Iowa.

\*University of Nebraska-Lincoln, Lincoln, NE 68588, and John Carroll University, Cleveland, OH 44118, respectively.

TABLE 1—PERCENT CORRECT ON TEL

Course/Items	Pre-Test	Post-Test	Change
Economics [1,499 cases]			
All Items (77)	44.9	52.4	7.5
Fundamental (20)	47.0	58.4	11.4
Microeconomics (20)	48.6	55.5	6.9
Macroeconomics (23)	41.0	46.5	5.5
International (12)	42.2	47.9	5.7
Consumer Economics [579 cases]			
All Items (77)	40.3	40.1	-0.2
Fundamental (20)	42.9	45.6	2.7
Microeconomics (20)	44.5	43.4	-1.1
Macroeconomics (23)	35.9	33.6	-2.3
International (12)	36.7	37.6	0.9
Social Studies			
with Economics [456 cases]			
All Items (77)	47.7	47.7	0.0
Fundamental (20)	49.4	50.4	1.0
Microeconomics (20)	53.4	52.0	-1.4
Macroeconomics (23)	42.2	42.6	0.4
International (12)	45.5	45.0	-0.5
Social Studies			
without Economics [497 cases]			
All Items (77)	37.4	36.9	-0.5
Fundamental (20)	39.7	41.9	2.2
Microeconomics (20)	40.9	39.6	-1.3
Macroeconomics (23)	33.4	31.9	-1.5
International (12)	35.0	32.9	-2.1

Note: Number of items is in parentheses.

merged test directly mirror those for each form.) The mean post-test level of economic literacy varies substantially for students in different courses. Students in social studies courses whose teacher did not include economics could correctly answer only 37 percent of the questions, or just 12 percent over a chance level on a four-option multiple choice test. The performance of students in consumer economics courses at 40 percent correct was only slightly better. Students in social studies courses where the teacher included economics score 48 percent correct, and economics students score 52 percent correct. Under the most liberal grading standards, and even considering the fact that the TEL was designed as a normed achievement test, these post-test scores would be classified as failing.

Subtest analysis was also conducted by calculating the mean percent correct for the post-test in each of the four major concept clusters defined in the *Framework*. The worst levels of performance are on macroeconomics and international economics items. For example, economics students score 47 percent correct on macroeconomic items and 48 percent correct on international items compared to 58 percent correct on funda-

mental items and 56 percent correct on microeconomic items. The results are similar for other courses. Students show about 6–10 percent less knowledge of macroeconomic and international concepts than they do of fundamental and microeconomic concepts. Weak performance in these key economic clusters is directly contributing to the failing grades on the overall test.

A more positive picture can be painted when the change from the pre- to post-test is examined, at least for students in the economics course. Economics students show a 7.5 percent improvement in the overall percent correct. Most of this gain comes from the increased understanding of fundamental concepts (+11 percent) versus the other concept clusters (+6–7 percent). In contrast, there is essentially no change in economic understanding in the other courses. Students in these courses show slight gains in understanding of fundamental items, but this gain is offset by slight declines in knowledge of microeconomic, macroeconomic, and international economic concepts. Consumer economics and social studies courses do not contribute much to economic literacy and are not effective substitutes for a separate course in economics as a means of increasing economic understanding.

Data are presented in Table 2 on the comparative performance of just the economics students on the economic concepts that form the four concept clusters. Concepts with the best scores (+60–75 percent correct) are, with the exception of unemployment, from the fundamental and microeconomic clusters and include: economic systems; economic institutions and incentives; money and exchange; and, supply and demand. Average performance (52–59 percent correct) is shown with such fundamental or microeconomic concepts as scarcity, opportunity cost/tradeoffs, productivity, markets and prices, competition and market structure, government, and with two macroeconomic concepts, GNP and aggregate demand. The lowest scores (+35–49 percent correct), with the exception of the low item score on market failure, are reserved exclusively for macroeconomic and the international items: aggregate supply; inflation;

TABLE 2—PERCENT CORRECT FOR ECONOMICS COURSE

Clusters/Concepts	Pre-Test	Post-Test	Change
Scarcity (3)	38.6	53.5	14.9
Opp. Cost./Tradeoffs (5)	42.6	52.2	9.6
Productivity (3)	45.7	52.3	6.6
Economic Systems (1)	62.8	75.0	12.2
Mon. Inst./Incent. (5)	51.1	63.4	12.3
Exc./Money/Interdep. (3)	52.2	64.5	12.2
Markets & Prices (2)	49.1	54.3	5.2
Supply & Demand (7)	52.2	61.0	8.8
Compet. & Struct. (4)	56.5	57.5	1.0
Income Distribution (3)	45.2	50.4	5.2
Market Failures (3)	34.2	42.6	8.4
Role of Government (3)	47.9	55.7	7.8
Gross Nat. Product (2)	52.1	59.0	6.9
Aggregate Supply (2)	38.8	45.4	6.6
Aggregate Demand (3)	47.0	54.9	7.9
Unemployment (2)	58.7	63.9	5.2
Inflation/Deflation (4)	32.8	35.3	2.5
Monetary Policy (5)	29.5	38.3	8.8
Fiscal Policy (5)	44.7	47.0	2.4
Comp. Adv./Trade (5)	46.2	51.8	5.6
Sal. Pay./Exc. Rates (4)	40.6	45.0	4.4
Economic Growth (3)	37.5	45.2	7.7

Note: Number of items is in parentheses.

monetary policy; fiscal policy; comparative advantage and trade barriers; balance of payments and exchange rates; and, economic growth.

## II. Regression Models and Results

Regression analysis of the overall *TEL* scores was conducted to identify factors that contributed to economic understanding. The analysis was necessary to control for the effects of any background variables that might not be accounted for in the item analysis. It could be claimed, for example, that one reason that students in an economics course performed better than students in other courses was because students in those courses were more intelligent or from higher income levels than the group of students in the other courses.

"Absolute level" and "absolute improvement" models (John Siegfried and Rendigs, 1979, p. 929) were specified for the analysis. The first model examines factors that contribute to the stock of economic understanding. It has been used in several previous national studies of high school economics (our 1982 article; Soper and Judith Brenneke, 1981; and, G. L. Bach and Saunders, 1965). The second model measures

TABLE 3—*TEL* REGRESSION RESULTS ( $N = 2,483$ )<sup>a</sup>

	Equation 1	Equation 2
Constant	-17.782 (8.025)	-17.399 (9.084)
<i>TELPRE</i>	-	0.536 (28.905) <sup>c</sup>
[20.34; 7.45]		
<i>IQ</i>	.295 (31.459) <sup>c</sup>	0.175 (19.194) <sup>c</sup>
[59.56; 15.12]		
<i>MALE</i>	1.345 (5.169) <sup>c</sup>	0.520 (2.296) <sup>b</sup>
[.51; .50]		
<i>SENIOR</i>	1.340 (4.430) <sup>c</sup>	0.599 (2.280) <sup>b</sup>
[.58; .49]		
<i>BLACK</i>	-1.633 (3.548) <sup>c</sup>	-0.871 (2.184) <sup>b</sup>
[.10; .30]		
<i>ECON</i>	4.128 (10.316) <sup>c</sup>	3.821 (11.043) <sup>c</sup>
[.54; .50]		
<i>CONECON</i>	-0.109 (0.219)	-0.121 (0.283)
[.15; .35]		
<i>SSECON</i>	2.435 (4.472) <sup>c</sup>	1.034 (2.185) <sup>b</sup>
[.12; .33]		
<i>TCOUR</i>	.639 (8.321) <sup>c</sup>	.408 (6.111) <sup>c</sup>
[4.23; 2.28]		
<i>DEEP</i>	1.633 (4.829) <sup>c</sup>	1.408 (4.815) <sup>c</sup>
[.43; .495]		
<i>SIZE</i>	4.645 (6.778) <sup>c</sup>	4.236 (7.150) <sup>c</sup>
[3.06; .23]		
<i>MINCOME</i>	2.287 (4.704) <sup>c</sup>	1.346 (3.195) <sup>c</sup>
[.76; .43]		
<i>HINCOME</i>	1.958 (3.145) <sup>c</sup>	0.334 (0.618)
[.14; .35]		
<i>SUBURB</i>	-0.329 (0.863)	-0.228 (0.691)
[.47; .50]		
<i>URBAN</i>	-1.037 (2.515) <sup>b</sup>	-0.897 (2.516) <sup>b</sup>
[.21; .41]		
<i>NEAST</i>	-1.619 (3.770) <sup>c</sup>	-0.218 (0.584)
[.14; .34]		
<i>SOUTH</i>	-0.829 (2.494) <sup>c</sup>	-1.042 (3.627) <sup>c</sup>
[.40; .49]		
<i>WEST</i>	-0.634 (1.105)	-0.278 (0.560)
[.12; .32]		
<i>R-square</i>	.488	.618
<i>SEE</i>	6.424	5.552

<sup>a</sup>Dependent variable = *TEL* [22.14; 8.98]. Note here and above: variable mean; standard deviation appears in square brackets. The absolute values of the *t*-statistics are shown in parentheses.

<sup>b</sup>Significant at the .05 level.

<sup>c</sup>Significant at the .01 level.

the flow of learning that occurs from a pre-test to a post-test by including the pre-test as a regressor. The availability of matched pre- and post-test data permitted us to estimate this model with a large, national sample of high school students for the first time in economic education research.

The variable labels, means, and standard deviations for the regressions are presented in the first column of Table 3. The *TEL* post-test score was the dependent variable in each equation. The *TELPRE* variable in

equation 2 was the pretest *TEL* score. Rather than duplicate the analysis for each form of the *TEL*, raw scores on form A of the *TEL* were equated to the raw scores on form B using a linear equating formula (William Angoff, 1984, p. 101). Each equation was estimated using the equated scores. Student *IQ* was estimated with scores on the *Quick Word Test* (E. F. Borgatta and R. J. Corsini, 1964) that was administered at the same time as the post-test *TEL*. Student data were also used to construct dummy variables (1 = yes; 0 = no) to capture the effects of class rank (*SENIOR*), gender (*MALE*), and race (*BLACK*).

Three factors were included in the model that have policy implications for economics instruction in senior high schools. First, course type differences were captured by three dummy variables, one for an economics course (*ECON*), one for a consumer economics course (*CONECON*), and one for a social studies course with economics (*SSECON*). The omitted category was a social studies course without economics instruction. Second, the influence of the economics human capital of the teacher was measured by the number of credit courses in economics that each student's teacher had taken (*TCOUR*). Third, information was collected on the degree of school district involvement in teacher training and curriculum development through the Developmental Economic Education Program (*DEEP*) sponsored by the Joint Council on Economic Education (John Maher, 1969). It was anticipated that students in *DEEP* districts that had implemented and sustained the program would outperform students in non-*DEEP* districts.

The remaining variables control for other background and environmental factors that might influence economic knowledge and learning. The estimated income of students in a class was represented by two dummy variables, one for high income (*HINCOME*) and one for middle income (*MINCOME*), with the excluded income class being low income. The size of the school (*SIZE*) in which the course was taught was included in the model, but transformed to common logs to correct for skewness in the distribution. The type of community in which the school

was located was controlled for by two dummies, one for an urban (*URBAN*), and one for a suburban (*SUBURB*) location, with the rural location serving as the excluded group. The census region for the school was captured by dummy variables representing the northeast region (*NEAST*), the southern region (*SOUTH*) and the western region (*WEST*), with the north central region serving as the comparison group.

The results from estimating equations by ordinary least square are provided in columns 2 and 3 of Table 3. All other things equal, the type of course a student takes has a significant effect on the level of economic knowledge in equation 1. Students who have completed an economics course score 4.1 points higher on the *TEL* than social studies students whose teachers *do not* include economics instruction in their courses. Social studies students whose teachers *do* include economics instruction in their courses score 2.4 points higher on the *TEL*. Students in a consumer economics course score about the same as students taking a social studies course without economics. These post-test rankings are similar to the results for the mean percent in Table 1.

As shown in equation 2, economics instruction also makes a contribution to the post-test score beyond that explained by *TELPRE* and the other variables. *ECON* students show a highly significant increase in knowledge by 3.8 points when compared with students taking a social studies course without economics instruction. *SSECON* students show a slight gain of 1 point on the *TEL* relative to students in the no-economics social studies course. Students in consumer economics courses learn no more economics than students taking a social studies course whose teacher does not include economics in the instruction. Obviously, the direct approach through a separate course makes the most significant contribution to economics learning, although the integration of economics in a social studies course may be somewhat helpful.

Teacher coursework in economics improves the economic knowledge of students. In equation 1, each college-credit economics course that a teacher has taken adds .64 of a point to the predicted *TEL* score. Moreover,



the more education a teacher has in economics, the more student learning of the subject increases. Even after accounting for the influence of the pre-test knowledge in equation 2, each course a teacher has taken still adds .41 of a point to student knowledge. These results provide further support for the value of teacher education in economics as a means of improving the economic literacy of high school students.

The *DEEP* variable is a significant predictor of economics achievement and contributes to gains in economic knowledge. Students in *DEEP* districts, which provide teacher in-service education in economics and which build economics into the curriculum, score 1.6 points higher on the *TEL* than students in non-*DEEP* districts. The contribution from *DEEP* does not disappear when the pre-test variable is included in equation 2 because there is still a 1.4 point difference in economic knowledge in favor of students in *DEEP* districts. The reasons for this effect are difficult to identify, but *DEEP* participation probably helps teachers by giving them access to curriculum materials, consulting assistance, and in-service education. These benefits, in turn, get incorporated into classroom instruction for students. *DEEP* is supposed to work that way and these results suggest that it does make a contribution to knowledge and learning.

The findings from the other variables will not be discussed because of space constraints and because most of these variables are not subject to policy changes. We now turn to the implications of these results for improving economic literacy in the nation's high schools and for teaching economics in college.

### III. Implications

Based on the test and regression analyses, we would recommend that several actions be taken in school districts to reduce the economic illiteracy of high school graduates. All high school students, whether job market or college bound, should take a separate course in economics because this course is the only reliable way to make significant gains in economic knowledge. There is some movement in this direction across the nation be-

cause at least 15 states now require a course in economics for high school graduation (Dennis Brennan, 1986, p. 20-1). Infusing economics into a social studies course may help, but it should not replace direct instruction in the subject; consumer economics may teach students about other topics that are not measured by the *TEL*, but that course does not add to economic knowledge.

The high school economics courses should devote more time to the study of macroeconomics and international economic concepts. Economics courses now do their best job in teaching students about fundamental economics and related concepts of scarcity, economic systems, economic institutions and incentives, and money and exchange. They even develop some understanding of the rudiments of supply and demand. However, high school economics students show an appalling amount of ignorance of *basic* concepts and relationships in macroeconomics and international economics which has nothing to do with theoretical disputes in the economics profession. Either economic concepts in these areas are not taught, or if they are taught, economics teachers do a poor job of providing instruction.

This last point raises another concern about the economic knowledge of teachers. The results clearly indicate that the more education in economics a teacher has, the better the students do and the higher the level of achievement. Teachers need to be encouraged to take more coursework in the everchanging field of economics if they are to stay current. One way to do this would be for a school district to make a stronger commitment to economic education through *DEEP*. Additional economic education provided to teachers through *DEEP* should also be supplemented with the creation of more curriculum materials and with more training in the use of the materials in the classroom. The preparation of new instructional materials on macroeconomics and international economics should increase knowledge of these topics.

Our findings suggest that significant improvements in the economic literacy of U.S. high school students will be made when students take an economics course, from teachers who have taken many economics

courses and who teach macroeconomics and international economics, and in a school district that has made a substantive commitment to economic education. Aside from personal, environmental, and demographic variables over which there is little control, these factors significantly influence the level of economic knowledge and increase economic learning. Until these changes are made, college instructors can safely assume that high school graduates who enter introductory economics courses are sadly deficient in their knowledge of basic economic concepts and relationships—a situation college instructors will have to correct. But the majority of high school graduates never go to college, and even when they do, they may not take a course in economics. Without solid education in high school economics, most adults will never have a chance of becoming literate in economics.

#### REFERENCES

- Angoff, William H., *Scales, Norms, and Equivalent Scores*, Princeton: Education Testing Service, 1984.
- Bach, G. L. and Saunders, Phillip, "Economic Education: Aspirations and Achievements," *American Economic Review*, June 1965, 55, 329–56.
- Brennan, Dennis C., *A Survey of State Mandates for Economics Instruction, 1985–86*, New York: Joint Council on Economic Education, 1986.
- Borgatta, E. F. and Corsini, R. J., *Quick Word Test Manual*, New York: Harcourt, Brace, and World, 1964.
- Hansen, W. Lee, "The State of Economic Literacy," in D. Wentworth et al., eds., *Perspectives on Economic Education*, New York: Joint Council on Economic Education, 1977, 61–79.
- Maher, John, "DEEP: Strengthening Economics in the Schools," *American Economic Review Proceedings*, May 1969, 59, 230–38.
- Saunders, Phillip et al., *A Framework for Teaching the Basic Concepts*, New York: Joint Council on Economic Education, 1984.
- Siegfried, John J. and Fels, Rendigs, "Research on Teaching College Economics: A Survey," *Journal of Economic Literature*, September 1979, 17, 923–69.
- Soper, John C. and Brenneke, Judith S., "The Test of Economic Literacy and an Evaluation of the DEEP System," *Journal of Economic Education*, Summer 1981, 12, 1–14.
- \_\_\_\_\_ and Walstad, William B., *The Test of Economic Literacy* (2nd ed.): *Examiner's Manual*, New York: Joint Council on Economic Education, 1987.
- Stigler, George J., "The Case, if Any, for Economic Education," *Journal of Economic Education*, Spring 1970, 1, 77–84.
- Walstad, William B. and Soper, John C., "A Model of Economics Learning in the High Schools," *Journal of Economic Education*, Winter 1982, 13, 40–54.

# Variables Affecting Success in Economic Education: Preliminary Findings from a New Data Base

By WILLIAM J. BAUMOL AND ROBERT J. HIGHSMITH\*

The output of the considerable effort expended on economic education in the schools of the United States is, of course, the resulting contribution to the students' degree of understanding of the workings of our economy and their pertinent reasoning ability. A comprehensive survey questioning both teachers and students, whose first phase has now been completed, permits a more extensive description of what has been achieved, of the means that have been used in the process and of the resources available for the purpose. Perhaps even more important, it makes possible an empirical analysis of the relationship between the inputs and the outputs—the methods and resources used and the achievements of the students. This article describes some of the main results obtained from a first analysis of these data.

Among the major results to date that emerge from this study are the following conclusions:

1) Students who receive formal training in economics at the senior level in high school, consisting of a minimum of three hours per week, understand some economic topics quite

well but have major gaps in their understanding of others.

2) Students share with their teachers many of the same goals for studying economics, but students believe that these goals are less important than teachers believe them to be.

3) Most economics students have not had any experience in studying economics prior to the senior level course.

4) Economics students believe that economics helps them to think more systematically about some kinds of issues they face, but not others.

5) Student attitudes toward their economics courses vary considerably, with 23 percent describing themselves as "liking economics a lot," 42 percent as "liking it a little," 16 percent as "unsure," 7 percent as "disliking it a little," and 6 percent as "disliking it a lot."

6) Senior students receiving formal training in economics constitute a fairly representative group in terms of Scholastic Aptitude Test scores but not in terms of the educational achievements of their parents.

7) Teachers giving instruction in economics at the high school level vary widely in the amount of their college training in economics and in their experience in teaching it.

8) The gender and ethnic backgrounds of economics teachers differ from those of economics students.

9) The topics in economics most frequently included by economics teachers in their courses were supply and demand, how market and prices work, and monetary and fiscal policy. Topics most neglected include balance of payments, how to interpret economic data, and measurement concepts.

10) The teachers felt that the instructional materials most helpful to them were newspapers, textbooks, and graphs and charts.

\*Professor of Economics, Princeton University, Princeton, NJ 08544 and New York University; and Director of Research, Joint Council on Economic Education, 432 Park Avenue South, New York, NY 10016, respectively. We are extremely grateful to the J. Howard Pew Freedom Trust, Inc. whose generous grant to the Joint Council on Economic Education and Princeton University made possible the research underlying this paper and the creation of the data base on economic education. The basic ideas for this work originated in the Joint Council, which provided supervision throughout the project. The data collection was carried out promptly and competently by Audrey McDonald. Invaluable suggestions and comments were provided by an advisory committee composed of William Becker, Indiana University; Marilyn Kourilsky, UCLA; Charles Plott, California Institute of Technology; and Sherwin Rosen, University of Chicago.

11) Teachers of economics believe that all teachers, themselves included, should be required to take considerably more courses in economics than they have taken.

All in all, the data base already provides us a clearer description of the instructors, of the facilities at their disposal and the methods they use. It also offers us a description of the students, the homes from which they come, and the coursework in which they participate. Above all, it can support and has already begun to support serious analysis indicating what influences make for increased effectiveness in the teaching of economics at the precollege level. It helps us in the search for readily reproducible approaches that can contribute to this effectiveness.

The remainder of this paper describes the nature of the data and the methods that were used to gather them. It indicates the methods that were used to analyze the data and reports descriptive statistics that underlie the conclusions that have just been summarized. It concludes with a description of the more advanced statistical work that is planned in the process of completion of the study.

### **I. Purpose of the Study**

The study reported here was intended to assemble a set of data that can contribute to future research on economic education, and to provide a first analysis of these data. The data that were collected pertained to twelfth-grade students and their teachers in public and private schools throughout the United States.

The goal of the study was to help to determine whether economic education is achieving the purposes for which it is intended, and to find out what circumstances contribute to the magnitude of that achievement.

The study proceeded on the premise that the central objectives of economic education are to equip students by the time they graduate from high school to understand enough about the discipline to make reasoned judgements on economic issues, and to give the student some understanding of

how the economy works. This, in turn, is intended to enable them to become more efficient decision makers and more responsible citizens, using reasoned analysis as a replacement for emotional judgements.

The study had four particular goals: 1) the provision of descriptive data on the current state of economic education which can serve as a benchmark against which future accomplishments can be measured; 2) the construction of a data base to be used by the Joint Council on Economic Education and others interested in the field as a basis for research; 3) analysis of the data to help in determining what new knowledge can be learned about how economics is taught in high schools, and what prior knowledge is reinforced through replication of the research on which it is based; and 4) a series of articles, some analytical, some more popular, with the latter intended to draw the attention of the public to the purposes and achievements of economic education.

To guide the study and facilitate its pursuit of the objectives, a set of general questions was formulated:

1) Does the study of economics contribute to the students' knowledge of pertinent facts?

2) Does this study contribute to students' understanding of the workings of the economy?

3) Does the study of economics contribute to the ability of students to reason analytically, especially about issues related to their future role as consumers, producers, workers, and voters?

4) What instructional materials, institutional arrangements, or other resources contribute most to effectiveness of a program or a course in economic education?

5) What characteristics and opinions of students most influence their ability to learn and their interest in economics?

6) What characteristics and attitudes of teachers contribute most to success in their teaching of economics?

7) What are the main barriers impeding economic education?

8) Do students' economic knowledge and attitudes in school districts participating in the Developmental Economic Education Program (DEEP) differ significantly from

that of students in non-DEEP districts? How and why?

9) In what ways and to what extent do the environmental conditions in the homes, school, and classroom from which students come influence learning of economics?

## II. The Survey

The data were collected by means of a survey conducted with the aid of four questionnaires. The students included 75 percent who had taken a course involving at least one hour of economics per week for 20 weeks, while 25 percent of the student respondents had received no formal economic education.

All of the students had taken a standardized examination, the standardized *Test of Economic Literacy (TEL)*, 2nd ed., Form B, see John Soper and William Walstad, 1986) that is intended to measure both the degree of their familiarity with economic facts and their facility in economic reasoning.<sup>1</sup> Their *TEL* scores were included in the data bank, thus permitting analysis of the relationship between student performance and the other data obtained from the survey.

In each case, students, teachers, schools, and districts were chosen so as to permit direct matching of the four sets of data. That is, it was known that student W had been taught economics by teacher X in school Y in district Z, so that if student W was included in the sample, then so was teacher X, school Y, and district Z.

After the questionnaires were distributed initially at each stage in the four-stage process (district, school, teacher, and student),

follow-up steps were undertaken to assure as large a response rate as was possible given the constraints of budget and time. These included follow-up mailings within three weeks of the initial mailing, telephone calls both from the surveying organization and, where appropriate, from the Joint Council. When repeated attempts at the district or school level met with failure, one of three reserve schools, selected randomly during sampling, that satisfied the same requirements as the school from which we were unable to achieve cooperation, was substituted.

The questionnaires were initially designed by us, then redesigned after consultation with the members of the steering committee and the group that carried out the survey. The questionnaire was then submitted to a pre-test group of 89 students and 2 teachers to minimize the incidence of badly worded questions and other problems in questionnaire design. All questions were also carefully prescreened to minimize redundancy and to ensure comprehensiveness. To avoid redundancy, no question was retained unless it was possible to describe in advance a particular use for the data that would emerge from it. To minimize gaps in the data, a very preliminary draft of this paper was prepared so that the nature of the pertinent information was spelled out explicitly.

It is planned to carry out a follow-up survey three years after completion of the first. Two purposes of the follow-up study will be to determine whether any indication of trends can be discerned, and to seek evidence of the effects of any changes in educational practices and other pertinent variables. Unfortunately, the brevity of the intervening period will naturally limit severely what can be learned on these two matters. More important will be the opportunity, made possible by the availability of students' names and addresses, to reinterview or re-question a subsample of the students in the original sample, in order to distinguish long-run achievements of the teaching programs from those that are merely transitory.

In addition, the follow-up study will make it possible to supplement the data acquired in the first round of the survey and to deal

<sup>1</sup>This is a test devised under the sponsorship of the Joint Council on Economic Education, and designed under the supervision of a committee that included James Tobin, Rendigs Fels, William Becker, and others. John C. Soper and William B. Walstad were the authors. The test deliberately includes some questions intended to test knowledge of relevant facts, some that test understanding of the workings of economic institutions, some that test knowledge of some economic theory and some questions that test the student's ability to use the sort of reasoning employed by economists.

with any lacunae that emerge in the meantime.

Let us now turn to a discussion of the results of the preliminary study of the accumulated information. Let us begin with the descriptive material and, in particular, with the data relating to the teachers.

### III. Who Teaches Economics in the High Schools?

Instructors who provide economics courses differ widely in number of years of their experience and relevant training.<sup>2</sup> The data indicate that the teachers' mean and median teaching years were 15.63 and 15, with a standard deviation of 8.55. The corresponding figures for the number of years they have taught economics were 7.63, 4, and 8.45.

Analogous figures for the number of credit hours of coursework in economics taken by the teachers during their undergraduate and graduate studies reveal that the number of relatively untrained teachers is quite large, with 25 percent of the high school economics teachers having accumulated *less than* 6 semester or quarter hours of course credits in the field and with an additional 29 percent having taken only 6 hours.

In addition, the data show that the average respondent had attended 3.26 in-service programs (for which no credit is given) lasting less than 1 day, 1.98 programs lasting 1–3 days, and 1.00 in-service programs lasting more than 3 days. Teachers had most recently participated in a college course or workshop two years ago on the average. Six percent of the teachers had majored in economics and the median teacher had received no preservice contact hours of training in techniques for teaching economics.

Altogether, the picture of the preparation of the high school teachers of economics that emerges from these and other data from the survey is that the majority of teachers during their undergraduate and graduate educations receive minimal training both in the content

of economics and in how to teach it. However, one in five teachers has taken more than 12 course hours in the subject, and most teachers have supplemented their university training with in-service training.

We turn next to teaching practices. The data indicate that, on the average, economics classes met 4.75 hours per week (st. dev. = 1.73), and 81 percent of the courses taken by seniors last for less than a full year. We see that few of the teachers are burdened with economics classes typically containing a substantial number of students. Moreover, most economics teachers teach only a single economics class to seniors during any given year.

Teachers were asked to rate from very important to unimportant, various goals for teaching an economics course. Teachers believe overwhelmingly that they are teaching students to understand the American economy in order to help them to make more intelligent decisions. Teachers are much less interested in treating alternative economic systems as a high priority goal.

Teachers also rated 22 economic topics taught in terms of percentage of time devoted to each. A considerable dispersion was revealed, but with relatively heavy emphasis upon tools of microeconomic analysis. Relatively, less emphasis is placed by economics teachers on tools of macroeconomic analysis and international trade.

The teachers were also asked various questions about the sorts of teaching aids and materials they would find most helpful and the levels of economics training they consider optimal for precollege instructors. Evidently, the teachers are most anxious to have newspapers, textbooks, and graphs and charts available to them, but not student newsmagazines, computer software, and help with test questions.

Asked their views on the proper amount of training in economics for an instructor in economics, the teachers responded that they feel their average current training is seriously inadequate. Thus, while for teachers in our sample the mean number of undergraduate economics courses taken was 3, with 6 percent having majored in the subject, according to the mean of the teachers' views the

<sup>2</sup>Readers desiring a copy of the tables on which these conclusions are based should contact Highsmith at the Joint Council.

appropriate number of courses is well over 3 courses, and 14 percent of the respondents considered an economics major to be desirable.

#### IV. The Students

We now turn to a description of the students who take a senior level economics course, as it emerges from the survey.<sup>3</sup>

First, let us describe their home environment. Economics students come from homes where the average level of education is significantly higher than the national average, and where parents take an interest in their studies, as evidenced by the fact that 75 percent inquire at least once per week about students' homework and one-half inquire almost every day.

The gender and ethnic backgrounds of economic students compared with other twelfth graders in the school and all students in the high school also were revealed in the survey. It was found that non-Hispanic whites are underrepresented in economics classes, as are females when compared with other relevant student groups in their high schools. Moreover, when the gender and ethnicity of students and teachers are compared, it becomes apparent that more economics teachers are white and fewer are black, Hispanic, or female relative to their students.

Economics students also hold jobs and attend, predominantly, public schools. Sixty-four percent of the students had jobs, working on average 13.5 hours per week (st. dev. = 12.25; median = 15 hours). Of the schools attended by students, 91 percent were public, and 9 percent were parochial and private (nondenominational).

Students varied considerably in academic accomplishments and other pertinent scores. Their cumulative grade point average in high school had a mean of 2.7, on a 4-point scale (median = 2.5; st. dev. = 1.48). For the 41

percent of respondents who indicated that they had taken the Scholastic Aptitude Test (SAT), economics students score higher on SATs than others in their school but very close to the same as others in their school district. A negligible percentage of students takes economics courses lasting a year or part of a year before their senior year in high school. Considerably more students, however, study economics prior to grade twelve as part of another class. Most students (72 percent) who take economics as seniors have never studied the subject before. Finally, by far the majority of seniors who study economics do so in courses lasting part of a year rather than a full year.

The students also varied widely in the interest they manifested in economics courses. Ranking them from 1 (= dislike a lot) to 5 (= like a lot), the mean score was 3.75 (st. dev. = 1.10). Fifty-six percent of the students indicated that they wanted to take another economics course.

When asked to indicate how much time they spend on economics homework in their current course, the seniors indicated that they spend 1–2 hours per week (st. dev. = 1.43) out of the 5 hours on homework they spend on average on all courses (st. dev. = 1.38). Turning to more objective indicators, the number of economics classes missed per student in a typical month varied from 0 to 10 days with a mean of 2.5 (st. dev. = 1.06).

The students' goals in studying economics were relatively heterogeneous. Learning practical skills such as how to balance a checkbook and how to make money are *very* important to more senior economics students than other goals. However, most of the other goals are deemed to be important, in equal measure. The one exception, understanding other economic systems, is believed to be the least important goal by the largest percentage of senior economics students.

Comparing the goals of an economics course selected by teachers and students reveals, with one exception, that students and teachers rank similarly the goals of teaching and learning economics. However, teachers see the goal within each rank as being considerably more important than students believe it to be. The exception is understanding

<sup>3</sup>In this section, unless explicitly noted otherwise, we leave out of consideration the students in our sample who had not taken any economics courses.

of practical skills like balancing a checkbook, which teachers rank relatively low and students rank as the most important goal.

Students varied widely, too, in their opinions regarding how the study of economics has helped them outside of school. Fifty-eight percent of the students believe that the study of economics helps them to think differently, at least to *some* extent, about the pros and cons of various jobs, whereas only 26 percent indicated that economics provided *some* help in using time wisely. In general, students received more help from the study of economics regarding major, long-term issues affecting their lives than they received regarding the day-to-day personal skills they require.

Finally on the *Test of Economic Literacy* (2nd edition, Form B), more than two-thirds of the economics students revealed an understanding of questions dealing with scarcity, markets and prices, economic institutions and incentives, and supply and demand. Fewer than one-third of the students correctly answered questions related to productivity, exchange-money and interdependence, market structures, aggregate demand, inflation and deflation and money policy.

### V. Conclusion

Space constraints did not permit us to describe our preliminary findings from the

surveys we also conducted of principals of schools and the superintendents of school districts attended by economics students in the data base. Time constraints have not yet permitted us to investigate the heart of the issue that most concerned us in the surveys—determination of the major influences on the effectiveness of teaching programs in high school economics.<sup>4</sup> Space and time have been sufficient, however, to alert the reader to the rich potential provided by the data base for conducting research in promising hypotheses as yet uninvestigated, and for replicating earlier work.

We invite you to participate with us in the challenging investigations we are about to undertake.

<sup>4</sup>A list of all variables for which data have been collected are available from the authors.

### REFERENCE

- Soper, John C. and Walstad, William B., *Test of Economic Literacy*, 2nd ed., New York: Joint Council on Economic Education, 1986.



# The Effects of Advanced Placement on College Introductory Economics Courses

By STEPHEN BUCKLES AND JOHN S. MORTON\*

The College Board with the assistance of the Educational Testing Service and a Development Committee of economists and high school teachers is creating an Advanced Placement Program in Economics. The course description and curriculum materials provided by both the College Board and the Joint Council on Economic Education will be completed in the spring of 1988. The first administration of the Advanced Placement examinations in economics is scheduled for the spring of 1989.

Advanced Placement examinations are designed to be taken by the most able high school seniors following course work in the relevant discipline. Scores are sent to colleges and universities. If the student performs sufficiently well, the appropriate department may recommend credit and/or advanced standing. The Advanced Placement Program currently provides course materials and examinations for twenty-six college introductory courses in fourteen fields. In 1987, 262,000 students from 7,776 secondary schools took 369,207 examinations and sent results to almost 2,200 colleges and universities. The three largest subject areas, English literature and composition, American history, and calculus, had a total of over 238,000 students taking Advanced Placement examinations. The purpose of this paper is to discuss the implications of the new economics examinations for high school and university economics courses.

## I. High School Economics Courses

A 1981 study, *The National Survey of Economic Education (NSEE)*, concluded that 87

percent of the nation's junior and senior high school students have the opportunity to take economics. Increasingly, an economics course is becoming compulsory. Twenty-seven states mandate economics as part of the secondary school curriculum. Fifteen states require at least a one-semester course in economics. New York and California recently mandated one-semester economics courses for all students.

Although more students are taking economics courses, often the economics would not be recognized as such by many economists. We find some clues in the names of the courses. In the *NSEE*, only 56 percent of the teachers who claimed to teach economics called their classes "economics." Twenty-seven percent called their classes "consumer education" and 13 percent were teaching "free enterprise" courses.

The norming data for the *Test of Economic Literacy* (John Soper and William Walstad, 1986) show that high school courses can be effective. The *TEL* measures student achievement on twenty-two concepts delineated in *A Framework for Teaching the Basic Concepts* (Phillip Saunders et al., 1984). Students completing a course called "economics" had a mean score of 23.57 out of 46. Students with a social studies course that includes economics had a mean score of 22.85 and students with a course called "consumer economics" had a mean score of 21.70. Students without such courses scored a mean of 18.37.

Many introductory college-level economics courses require sophomore standing as a prerequisite. Even without such a prerequisite, students are often advised to wait until they are sophomores to begin the study of economics. The likely rationale for such requirements and recommendations is that freshmen lack the maturity and background to be able to understand the content of the introductory course. Implicit in the decision

\*Department of Economics, University of Missouri, Columbia, MO 65211, and Office of Economic Education, Governors State University, University Park, IL 60466, respectively.

to create Advanced Placement examinations in economics is the assumption that not only can freshmen learn college-level introductory economics, but some high school seniors are capable of doing so.

Studies comparing performance of freshmen and sophomores in introductory courses have found that there is no significant difference in scores on standardized tests. There is limited evidence that a significant minority of high school students can learn the economics taught in the college principles course.

We collected background and test score data describing a small sample of students at a high school which requires a year of economics for graduation. About one-third of the students take an honors course, the content of which is similar to that of the typical college principles course. College-level textbooks and supplementary material are used. Because consumer economics is mandated in the state, some of the content must cover consumer topics such as careers, money management, and investing. The cost of implementing the state's consumer education mandate is the elimination of extensive discussions of factor markets and some macroeconomic concepts.

The students in honors economics courses are better than average students. In 1986-87, the students in this study had a mean score of 26.8 on the ACT, 539 on the SAT Verbal, and 614 on the SAT Math. These scores were considerably higher than the national, state, and school scores. The national mean scores, for example, were 18.7 on the ACT, 430 on the SAT Verbal, and 476 on the SAT Math.

The students enrolled in the honors courses showed significant gains and had much better than average scores on the *TEL* and the microeconomics and macroeconomics versions of the *Test of Understanding College Economics (TUCE)*. On the *TEL*, the students gained 9.13 points or 214 percent of the national gain of 4.26 points. Within this small sample, all students scored at the end of the course above the mean of the national sample for high school students taking economics. The mean macro-*TUCE* score was in the sixty-eighth percentile of college stu-

dents completing an introductory course. The mean micro-*TUCE* score was at the seventy-seventh percentile. Thirty-nine of the 52 students scored above the mean of the national sample for macro-*TUCE* for students completing a macroeconomics course. Forty-three of the 52 scored above the national mean on the micro-*TUCE*. These data lead us to a conclusion that the level of economic understanding of very good students who have taken a high school course in economics can be compared to the economic knowledge of the better students after a college course in economics.

In the national norming data for the *TEL*, 32 percent of the scores of high school students with an economics course scored above the minimum score of this small sample. If the correlations we found between *TEL* and *TUCE* scores hold for the national sample, the level of economic understanding of approximately 25 percent of high school seniors who complete economics now is roughly equivalent to the level of understanding of the upper half of college students following an introductory course.

Although this is a limited sample, it could be hypothesized that higher-than-average ability high school students can learn economics at the college level. Because our sample program enrolls a larger percentage of students than a typical Advanced Placement course, and because several weeks are taken up with consumer economics, test results on an Advanced Placement course might be higher than they were in our sample. Most college courses have 45 contact hours each semester. High school courses have the advantage of having 90 contact hours each semester. This additional time should increase the probability that high school students can learn college level introductory economics.

Studies in other disciplines that have Advanced Placement examinations show that students with Advanced Placement credit learn as much as students who complete an introductory course while in college. Warren Willingham and Margaret Morris (1986) summarized a large number of studies which compare scores on the Advanced Placement exams by high school students and by

college students completing introductory courses. After adjusting for differences in abilities, the scores have been found to be comparable. Patricia Casserly (1986) reported the results of evaluations that show that Advanced Placement students do as well or better in upper-level courses in the Advanced Placement discipline than students who take their introductory courses in college. If these results are true for high school students taking courses in such subjects as biology, calculus, chemistry, and physics, we see little reason that the same result should not occur in economics.

## II. Effects on High School Economics

Empirical evidence on the effect of the Advanced Placement Program on high school and college economics will be available once we have experience with the program. Hypothesized effects have been suggested by the original Economics Advanced Placement Task Force and by experiences in other disciplines. The economics program should improve the quantity and quality of economics currently taught to high school students. The Advanced Placement exam should help standardize the economics curriculum offered to high-ability students and improve the quality of the textbooks used in high school courses. The typical high school economics course now lasts one semester and is often a mixture of economics, consumer education, and personal finance. The incentive of college credit for high school economics should encourage schools to offer a one-year course designed around the concepts in the college principles course. The result should be an increase in the economics concepts learned. In addition, with the added incentive of possible college credit, more of the high-ability students should be attracted to the courses.

Surveys were sent to the 208 secondary schools with the largest number of Advanced Placement candidates in American history. One hundred and ten schools responded. Sixty-six percent currently offer a one-semester micro- and macroeconomics course, 5 percent have a one-semester macroeconomics course, and 15 percent have two-semester courses. The other 14 percent offered no

semester-long economics courses. Seventy-eight percent of the secondary schools reported that they have staff qualified to teach an Advanced Placement course in economics. Fifty-nine percent of the schools stated that they were interested in offering an Advanced Placement course; the others were unsure. The College Board is projecting that over 10,000 students will take the examinations in the first year.

Walstad and Soper (1987), among others, have found the not-surprising result that the number of economics courses high school teachers have taken does positively and significantly affect student learning of economics. In Walstad and Soper's national sample, teachers of social studies and economics had taken an average of 4.4 courses in economics. Eighty-four percent of the sample were teaching economics, consumer economics, or social studies with economics included. This is a significant increase in the formal economics background of high school teachers from that found in research in the 1960's and 1970's. Whether 4.4 courses in economics is sufficient preparation to teach an Advanced Placement course in economics is an empirical question that we should be able to answer in the future. In any case, many high schools may not have teachers qualified to teach college-level introductory courses. The Joint Council on Economic Education with the assistance of a grant from the Alfred P. Sloan Foundation is developing an extensive set of materials to help train Advanced Placement teachers. Schools which normally offer Advanced Placement courses may encourage teachers to enroll in university inservice courses. Because of the status of teaching Advanced Placement courses, teachers have an additional incentive to learn more economics. Increased inservice economics education should improve the quality of instruction. In fact, the greatest benefit of an Advanced Placement economics test may not be for the students enrolled in Advanced Placement courses. The Advanced Placement Program will provide an opportunity for universities to train a corps of master economics teachers who will teach not only the Advanced Placement course, but also the regular courses. More

highly trained teachers would lower the risk of ersatz economics being taught in the high schools. In addition, Advanced Placement teachers will likely become advisors for the entire K-12 economics curriculum and provide guidance on the appropriate content of precollege economics. Because Advanced Placement economics teachers would have a self-interest in promoting economics, there would be more advocacy of economics in social studies departments. The creation of the exam itself will raise the prestige of economics within social studies departments, where there are currently exams in American and European history, and political science.

An Advanced Placement economics test presents potential costs as well as benefits. The curriculum of the typical college-level principles of economics course is more technical than the typical high school course. If the much-criticized encyclopedic type of course is simply transferred to the high school level, the result may be that little additional economics is learned. If the courses do not demonstrate the excitement and relevance of economics, fewer students may be attracted to the study of economics at the college level.

Another potential cost is that a full-year course may inappropriately crowd out other subjects in the high school curriculum. To provide choices for high schools (and easier interpretation of the results by college departments), two one-semester courses are being created. Schools will thus have the option of offering either a one-semester or a full-year course. In the analysis of our high school sample, we found that verbal skills as measured by SAT verbal scores significantly and positively affected performance on the macro-*TUCE*. Math skills as measured by SAT math scores significantly and positively affected performance on the micro-*TUCE*. These results, once verified in larger samples, suggest that a high school student able to take only a micro or a macro course might be guided into the appropriate course based on relative verbal and math SAT scores.

### III. Effects on College Economics

The existence of Advanced Placement examinations and courses at the high school

level should also have consequences for college-level instruction. If a large number of students take the Advanced Placement examinations and receive college credit, the better students at those institutions accepting Advanced Placement credit will not be enrolled in introductory courses. At those institutions not accepting credit, the repetition of an introductory economics course may be unattractive to some of those better students.

To the extent that the Advanced Placement program has a positive effect upon the level of economic understanding of all graduating high school students, the college-level introductory course may be able to accomplish more. However, if the effect on high school students is mixed in that only some students know significantly more, the variation in the level of economic understanding of students taking introductory courses may be increased. This would make teaching the traditional introductory course even more difficult than it is.

Few students enter college wanting to major in economics. Most economics students make that decision during or after the introductory course. Early significant exposure to economics may advance that decision for some students. There is evidence that students taking Advanced Placement exams are more likely to major in those subjects. Willingham and Morris found that Advanced Placement students were more likely to major in the field in which they took Advanced Placement exams. Students were twice to five times as likely to major when compared to non-Advanced Placement students with identical backgrounds. If the same holds true for economics Advanced Placement students, we may see an increase in the number of our majors. We may see enrollment in our intermediate-level courses increase beyond the change caused by more majors. Willingham and Morris found that even among nonmajors, Advanced Placement students enrolled in significantly more courses within the discipline than did other students.

With more and better prepared students coming into introductory and intermediate level courses, acceptance and support of economics education activities in elementary and

secondary schools may also become more widespread throughout the economics profession.

#### IV. Content of the Examinations

The Advanced Placement examinations are designed so that cognitive level and content coverage reflects what the Advanced Placement Development Committee believes to be the typical college introductory course. The coverage of the examinations combined with course descriptions and teachers guides that reflect that coverage should result in high school courses that are similar to the typical college introductory course.

The Advanced Placement economics exams consist of one examination in microeconomics and one in macroeconomics, each including one hour of testing using approximately 40 multiple choice questions and 30 minutes devoted to essay questions. The content specifications of the examinations are based on the results of a survey undertaken by Educational Testing Service and the initial Advanced Placement Task Force. The survey was sent to the 200 colleges and universities enrolling the largest numbers of Advanced Placement students. Ninety-nine institutions responded. The sample of responding institutions represents a broad spectrum of four-year higher education institutions. Forty-two of the institutions are research universities according to the classification scheme of the Carnegie Foundation for the Advancement of Teaching; 16 are doctorate-granting universities; 16 are comprehensive universities and colleges; 19 are liberal arts colleges; and 2 are in other categories. Thirty-seven percent of the responding institutions offer two-semester and one-semester alternatives. Forty-eight percent offer only a two-semester sequence. Of those institutions with two-semester courses, 24 percent require or recommend that students take macroeconomics first; 20 percent require or recommend that students take microeconomics first. The others have no recommended sequence.

The survey results identified the allocation of class time to specific content areas. The summary statistics for the two-semester sequence were as follows: 7.5 percent of

class time for basic economic concepts; 39.8 percent for macroeconomic concepts; 42.7 percent for microeconomic concepts; 6.2 percent for international economics; and 3.8 percent for applied areas and special topics. The Advanced Placement exams are designed to reflect the allocation of class time to specific content areas within each of these categories.

The results of the survey do provide some evidence on whether the typical course is an encyclopedic one. Less than one class period on average is devoted to the following topics: public goods; government regulation of consumer and worker safety; tariffs and quotas; financial markets; forms of business organization; growth theory; the environment; energy; agriculture; developing economies; and comparative systems. In most of these areas, fewer than a third of the courses even include the topics.

#### V. Conclusion

If our hypotheses and the evidence from other disciplines are applicable to economics, we can reach several conclusions about the effects of the Advanced Placement examinations. Teachers may need additional training. High school courses in economics should face increasing demand. The quality of high school economics courses should improve and have significant external benefits on the quality of economics taught in other courses and grade levels. The Advanced Placement Program should have a positive effect on the overall level of understanding of economics among the nation's high school graduates.

College and university economics departments may have to make adjustments in their introductory courses. Economics departments should experience an increase in enrollments and in majors, particularly among the better students. The Advanced Placement Program in economics should, on balance, benefit the profession.

#### REFERENCES

- Casserly, Patricia Lund, "Advanced Placement Revisited," College Board Report No. 86-6, New York: College Entrance Ex-

- amination Board, 1986.
- Saunders, Phillip, *Revised Test of Understanding in College Economics*, New York: Joint Council on Economic Education, 1981.
- \_\_\_\_\_. et. al., *A Framework for Teaching the Basic Concepts*, New York: Joint Council on Economic Education, 1984.
- Soper, John C. and Walstad, William B., *Test of Economic Literacy*, New York: Joint Council on Economic Education, 1986.
- Walstad, William B. and Soper, John C., "What is High School Economics? Factors Contributing to Student Achievement and Attitudes," unpublished, 1987.
- Willingham, Warren W. and Morris, Margaret, "Four Years Later: A Longitudinal Study of Advanced Placement Students in College," College Board Report No. 86-2, New York: College Entrance Examination Board, 1986.
- The National Survey of Economic Education, 1981*, New York: Playback Associates, 1981.

## Fluctuations in Equilibrium Unemployment

By ROBERT E. HALL\*

Macroeconomists tend to play down the role of fluctuations in product demand in their accounts of the movements of employment and unemployment. In the reigning mode of thought, shifts in demand account only for transitory departures of the unemployment rate from its natural level. The more persistent movements of unemployment, which account for the bulk of its variance around a long-term mean, are assigned to changes in the natural rate itself. Current thinking about the natural rate puts its emphasis entirely on the supply side of the labor market. The recent decline of the U.S. unemployment rate to below 6 percent is seen as the consequence of the declining importance of young people in the labor force, just as the bulge in unemployment in the early 1970's was blamed on the baby-boom generation.

I offer the contrary view that shifts in product demand generate movements in employment and unemployment for more than a transitory period of disequilibrium. In this view, the determination of employment and unemployment takes place in an equilibrium setting where demand is on an equal footing with supply. However, the equilibrium is not necessarily the one achieved by perfectly competitive markets and strict profit maximization by the owners of firms. Both the level of unemployment and the amplitude of its fluctuations may be aggravated by certain features of the economy. In general, these features generate substantial wage differentials between vestibule workers and perma-

nent workers. For example, agency problems may prevent the owners of firms from operating their personnel policies so as to maximize the value accruing to owners. Instead, employees have an important role in hiring and firing, and are able to raise the value of long-term employment above the value of alternative activities. Or, limitations in worker-discipline devices may make it optimal for compensation to rise with experience faster than does productivity.

When permanent jobs pay substantially more than entry-level jobs, the rate of promotion is a critical determinant of the economic value derived by a worker taking a job at the entry level. In the model of this paper, movements of promotion rates resulting from variations in the rate of growth of permanent employment generate important variations in the shadow wage in the entry-level labor market, even though the actual cash wage is hardly variable. Changes in equilibrium unemployment are the result of positive elasticities of labor supply with respect to the shadow wage. To put it another way, unemployment rises in youth labor markets when opportunities for advancement to high-paying permanent jobs decline, even though there is no change in the immediate cash wage.

### I. Unemployment and Labor Supply

This paper focuses on the young adult labor market, involving workers aged from 18 to 30 when the transition from education and other nonwork activities into serious full-time work takes place. A good deal of unemployment arises from this group, and much of the growth of equilibrium unemployment in the past two decades has been concentrated in these ages. I do not attempt to deal with unemployment among younger

<sup>†</sup>*Discussants:* Olivier J. Blanchard, MIT; James Medoff, Harvard University; Robert Topel, University of Chicago.

\*Hoover Institution, Stanford University, Stanford, CA 94305.

teenagers seeking summer or after-school employment, or with the much lower unemployment among workers over age 30.

Much unemployment among young adults arises from their participation in activities other than career-related employment. The U.S. unemployment survey (the *Current Population Survey*) does not count an individual as unemployed if even one hour of work occurred in the survey week. On the other hand, people who are in school, spending time at home, or taking some time off are quite likely to be recorded as unemployed—any job-seeking activity, even just looking in the want ads, in the four weeks before the survey will classify a nonworker as unemployed. Hence, it seems reasonable to view unemployment in this group as primarily the complement of permanent employment. A substantial and stable fraction of young adults not started on career paths will be counted as unemployed. An explanation of variations in unemployment is virtually the same thing as an explanation of variations in employment. Times of high unemployment are times when a smaller fraction of young adults are started on career ladders. The fact that the great bulk of unemployment arises from individuals who work relatively little (Kim Clark and Lawrence Summers, 1979) supports this view of unemployment.

## II. A Model of Labor Supply Based on the Timing of Entry

The point of the simple model of this section is to show how the shadow or effective wage that drives labor supply takes account of the likelihood of promotion to higher-wage jobs. The model shows that promotion probabilities can be an important determinant of labor supply. The model focuses on the timing of entry to the labor force as the most important dimension of labor supply. To keep the model simple, I assume that activities before entry do not involve employment in the formal labor market, though in reality the activities may include part-time or episodic work. The model has the implication that entry occurs only once, which is another reasonable simplification, although it is not universally true.

Consider a young worker whose preferences about work are ordered by the utility function,

$$(1) \quad \int_t^\infty e^{-rs} y(s, t) ds + v(t).$$

Here  $t$  is the time of entry to the labor force,  $r$  is the discount rate,  $y(s, t)$  is expected earnings at time  $s$  given entry at time  $t$ , and  $v(t)$  is the utility derived from activities preceding entry.  $v(t)$  is an increasing, differentiable function satisfying  $v''/v' < -r$ . The first-order condition for the optimal time of entry is

$$(2) \quad -e^{-rt} y(t, t) + \int_t^\infty e^{-rs} \frac{\partial y(s, t)}{\partial t} ds + v'(t) = 0.$$

Define  $q(t)$  as the shadow value of employment or the effective wage:

$$(3) \quad q(t) = y(t, t) - \int_t^\infty e^{-r(s-t)} \frac{\partial y(s, t)}{\partial t} ds.$$

The effective wage comprises the immediate cash earnings of a new entrant,  $y(t, t)$ , plus the present discounted value of the subsequent advantage of earlier entry,  $-\partial y(s, t)/\partial t$ . If entering sooner makes promotion to a higher-wage job more likely to occur sooner, that is an extra value beyond the immediate cash wage.

In the steady state, where  $q(t)$  is constant over time, the first-order condition can be written compactly as

$$(4) \quad e^{rt} v'(t) = q.$$

Let the function  $\theta(q, r)$  be the optimal time of entry:

$$(5) \quad e^{r\theta(q, r)} v'(\theta(q, r)) = q.$$

Under the assumptions stated, it is apparent that a higher effective wage brings earlier entry:  $\partial \theta(q, r)/\partial q < 0$ .



Now consider the labor force participation rate over the range  $0 \leq t \leq T$ , corresponding to ages 18 to 30. All those aged below  $\theta(r, q)$  will be out of the labor force and those above will be in the labor force. The labor force participation rate or labor supply function will be

$$(6) \quad n(q, r) = 1 - (\theta(q, r)/T).$$

Labor supply is an increasing function of the effective wage:  $\partial n(q, r)/\partial q > 0$ . A higher effective wage stimulates early entry and thus more total work from the age group. Unemployment is correspondingly lower. These changes can occur without any change in the current cash wage if the promotion rate increases.

As an example of the way that promotion rates influence the effective wage, consider the case where a worker starts at the entry wage  $w$  and faces a constant hazard,  $\pi$ , of promotion to a long-term job paying  $z$ . That is, after working for a length of time,  $\tau$ , the probability of being in the better job is  $1 - e^{-\pi\tau}$ . Then equation (3) shows that the effective wage is

$$(7) \quad q = \frac{r}{r + \pi} w + \frac{\pi}{r + \pi} z.$$

Figure 1 shows the effective wage as a function of the promotion rate for the case where  $r = 0.2$ ,  $w = \$5$ , and  $z = \$15$ . There is a substantial difference between the effective earnings of the entry job with little probability of promotion ( $\pi = 0$ ) and with a 30 percent annual probability of promotion ( $\pi = 0.36$ ).

It remains to establish a relation between promotion rates and the growth of a firm or an industry. Is it plausible that promotion rates fluctuate enough to cause important variations in the effective wage, given what is known about changes in growth rates of the sectors of the U.S. economy? Suppose that there are  $N_1$  vestibule workers and  $N_2$  permanent workers in a sector and let  $\alpha$  be the ratio of vestibule to permanent employment. Suppose that  $N_1$  and  $N_2$  and hence total employment are growing at the same rate,  $g$ . The flow of newly promoted workers is  $\dot{N}_2$ . The number of vestibule workers is

Effective wage

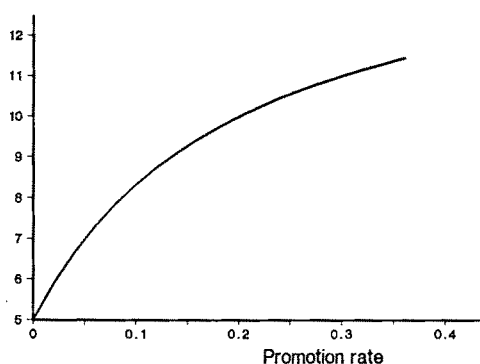


FIGURE 1. EFFECTIVE WAGE AS A FUNCTION OF THE PROMOTION RATE

$N_1$ . Hence the promotion rate is

$$(8) \quad \pi = \frac{\dot{N}_2}{N_1} = \frac{\dot{N}_2}{N_2} \frac{N_2}{N_1} = \frac{g}{\alpha}.$$

If vestibule workers are, say, one-tenth as numerous as permanent workers, then variations in the promotion rate are ten times as large as variations in the growth rate. Fluctuations of 1 or 2 percentage points in employment growth correspond to fluctuations of 10 or 20 percentage points in the promotion rate.

### III. Sources of Wage Differentials and Fluctuations in Promotion Rates

Economists have devoted great effort to studying and explaining wage differentials between beginning and senior workers. Not all explanations are consistent with a view that changes in promotion rates are a source of variations in the effective wage for starting workers. For example, there may be unobserved differences in productivity of workers at the time of hire. During the probationary period, firms accumulate information about productivity. They promote the workers who have a comparative advantage in working permanently for the firm. In that case, it is not easy to explain why promotion rates should vary with rates of growth. And even if promotion becomes more likely when

employment growth is high, the productivity of the promoted workers is likely to fall in relation to the entry-level workers, in which case the wage differential will probably shrink. An optimal promotion model seems unpromising as a way to explain important fluctuations in effective entry-level wages.

#### A. *Implicit Bonding*

One model, widely discussed in the literature on earnings profiles by seniority, explains wage differentials as a way to provide an appropriate penalty for workers who depart from jobs. The loss of future wages discourages workers from quitting and dissipating firm-specific human capital. The loss also makes threat of discharge an effective disciplinary device. The discharged senior worker must suffer some loss in earnings for the threat to be meaningful. In efficiency wage models, the loss occurs because the worker undergoes a period of unemployment before finding a new job. The alternative suggested here is that the more important wage loss is the period spent as a low-paid entry-level worker.

#### B. *Agency Problems in the Firm*

Another reason that firms might pay their permanent workers more than the market wage is that wage setting is imperfectly controlled by the owners of the firm. To some extent, employment decisions are made by employees, not by owners. The result is to divert some of the profit of the firm to employee-managers and away from the shareholders. The essence of the problem is the owners' inability to monitor their employees' efforts.

If the owners of a firm cannot observe their employees' efforts, it is desirable that they provide finance on a noncontingent basis. When the employee has no equity interest and the owner has a 100 percent equity interest, the employee is playing the game with someone else's money. Absent close monitoring, the performance of the employee is unlikely to be efficient. The answer is to make the employee play with his own money, by making the return paid to the owner be unrelated to the actual perfor-

mance of the firm. Thus employees are held to an exogenous standard of return on equity, or are required to pay a predetermined dividend. Whatever they make beyond that standard they are allowed to keep for themselves. The result will be earnings premiums in jobs where performance has exceeded the expectation formed when the system was established.

Now if one or a few of the workers are designated managers and can run the whole firm efficiently, the agency problem just set forth would have little impact on hiring and firing of the bulk of the work force. The managers would maximize profit by equating the marginal revenue product of labor to the market wage. But if a few managers could turn that trick, so could the owners. The agency problem exists throughout the firm. High-level managers are unable to monitor middle managers completely, who in turn are unable to monitor the heads of individual units. A firm is actually an association of worker-partners, none of whom is fully able to monitor the performance of the other worker-partners.

Figure 2 shows the standard analysis of the worker-managed firm. If all workers share equally in profit, and no worker outside the firm has a voice in employment decisions, employment will be set at the point  $L^*$  that maximizes the average revenue product of labor. At this point the marginal revenue product equals the average. The owner would choose the higher level of employment,  $\hat{L}$ , where the marginal revenue product of labor equals the market wage.

In practice, workers do not share equally in profits. Entry-level workers have nowhere near the claim of their senior brethren. By taking on workers in the vestibule, paid low wages, the partners holding the premium jobs can raise their own earnings above the maximum of the average revenue product of labor. They spread the profit created by all workers over the smaller number of worker-partners. However, if seniority is the only basis for excluding new hires from partnership, then the vestibule workers have to be let go before they become partners. Hence continual turnover of vestibule workers occurs. The turnover has the added benefit of permitting the partners to select only the

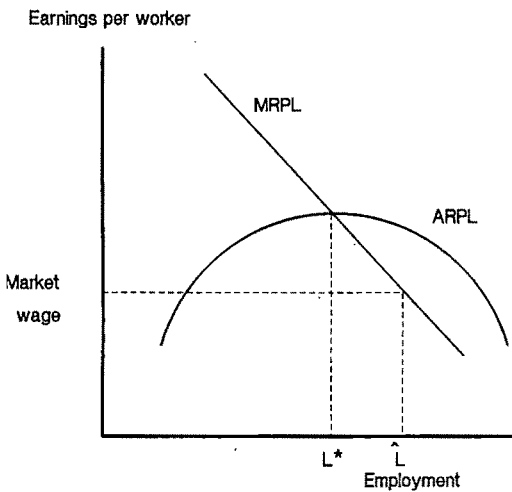


FIGURE 2. THE WORKER-MANAGED FIRM

most productive workers for partnership in case demand expands and it becomes attractive to add a long-term worker.

Because high-paid partners are recruited from among the vestibule workers, the cash wage is not the only attraction of vestibule employment. The total value is the cash wage plus the value associated with the probability of promotion. The value of the probability of promotion is the wedge separating the wage as perceived by the partnership and the wage as perceived by the worker. Promotion is much more likely when the firm is growing; the probability of promotion drops to close to zero when the number of partners is above the level that maximizes their average revenue product. The value of employment fluctuates much more than does the cash wage. Or, to put it the other way around, the firm finds it necessary to pay almost as high a cash wage in bad times as in good, because in bad times it cannot offer as high a chance of promotion. This situation is a consequence of the agency problem and the partnership rule that confers an instant capital gain to a new partner.

#### IV. Empirical Implications and Tests

In the labor market described above, there is no fixed natural rate of unemployment. Measured unemployment changes when incentives change. In times when vestibule jobs

offer high effective wages because promotion rates are high, young workers are attracted to those jobs away from other activities. Having a job is virtually a bar to classification as unemployed. Hence the move into jobs is accompanied by a decline in measured unemployment. In slack times, vestibule jobs are significantly less attractive even though they may pay the same cash wage. Unemployment is correspondingly higher.

The rate of growth of premium jobs is anything but a given constant over time. Consequently, the equilibrium rate of unemployment is variable over time. Periods of rapid growth of those industries with premium jobs will be periods of low equilibrium unemployment. When a shock occurs that is adverse to those industries, equilibrium unemployment can rise and remain high for many years. The theory permits highly persistent movements of unemployment; it does not require the return of unemployment to a constant normal level after a transient period of disequilibrium.

Measuring the growth rate of high-wage jobs is a relatively difficult task. Ideally, jobs paying premiums over wages elsewhere for workers with similar characteristics could be identified in cross-section research, and then the growth of employment measured in time-series data. What I have done to date is to observe that wage premiums are frequently found to be substantial in manufacturing in comparison to most other sectors. I have used the rate of growth of manufacturing employment as a rough measure of the opportunities available for promotion to advantageous long-term jobs. The postwar period has seen important changes in manufacturing employment growth over periods considerably longer than the business cycle. Growth was moderate in the 1950's, strong in the 1960's, slightly negative from 1970 until the early 1980's, and quite negative through the mid-1980's. In the work presented here, I take growth rates over 5-year periods to filter out some of the transitory business cycle effects.

With regard to unemployment, it appears possible to isolate the type of unemployment considered in this paper by taking the difference between total unemployment and insured unemployment. Insured unemploy-

ment is close to a trendless business cycle indicator; the notable increase in unemployment in the 1970's was almost entirely outside of insured unemployment (Gary Burtless, 1983). Uninsured unemployment is close to the concept of unemployment considered in the model of Section I, where unemployment arises among people who have not yet gained much job experience. The state-federal unemployment insurance system requires an extended period of employment in order to become eligible for very many weeks of unemployment compensation. As a result, few of the unemployed who have not yet started career employment receive benefits, even if their unemployment was preceded by a spell of employment.

Figure 3 shows the scatter diagram of manufacturing employment growth on the horizontal axis and uninsured unemployment on the vertical axis, measured as 5-year averages. There is a remarkably strong negative relation. The flow of job opportunities in manufacturing has declined almost continuously over the postwar period, except for the strong spurt of growth in the mid-1960's. The decline in manufacturing opportunities has coincided with an upward movement in the type of unemployment associated with delayed entry into serious career work.

Obviously the evidence in Figure 3 leaves many questions unanswered. Was there a similar decline in effective starting wages in sectors other than manufacturing? Is there more direct evidence other than uninsured unemployment about the changing behavior of young adults? But the evidence does suggest that changes on the demand side of the labor market have an important role in explaining changes over time in the natural rate of unemployment.

Summers (1986) considers a wide variety of evidence bearing on the issues examined in this paper. First, he shows quite convincingly that the upward trend in unemployment is unlikely to be the result of compositional shifts in the labor force. Second, he notes that unemployment has risen most among people who were previously employed, rather than new entrants to the labor force. Third, he notes that insured unemployment has not risen along with total un-

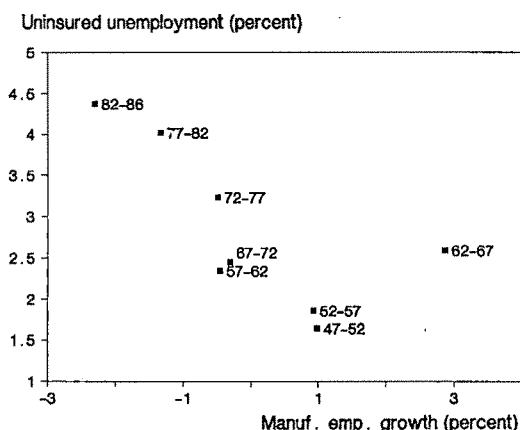


FIGURE 3. EMPLOYMENT GROWTH AND UNEMPLOYMENT

employment, a fact he considers a mystery. He is critical of the view taken here that changes in unemployment among those not yet fully committed to work are an important part of the explanation of the longer-term movements of unemployment. His primary evidence is that the rise in unemployment has not been concentrated among unmarried men. Ultimately, he concludes in favor of the view that more frequent and longer spells of unemployment between jobs are the principal source of higher unemployment. However, he does not develop a model in which this change is interpretable as an equilibrium. Moreover, Summers' inability to incorporate the rise in uninsured unemployment into his overall view is a significant shortcoming, in my view.

## V. Conclusions

When firms pay wage premiums to long-term workers, unemployment among young people is sensitive to the rate of growth of employment. The natural rate of unemployment is not fixed by supply considerations alone. Strong growth attracts entry-level workers because of the good prospects for promotion to long-term employment. In weak conditions, that factor is absent or attenuated and firms cannot take advantage of slackness by offering significantly lower cash wages. To put it differently, jobs that are attractive when demand is strong be-

cause they are on the first rung of a job ladder become dead-end jobs in slack times.

The effect of a decline in growth is anything but transient in this model. The period during which promotion to long-term jobs is closed off could easily be five years or a decade. The time pattern of regional differentials in unemployment fits this model reasonably well—they tend to come and go over near-decade periods. Similarly, as long-term demand shifts for durables occur, corresponding changes in employment and unemployment in that sector will occur.

## REFERENCES

- Burtless, Gary, "Why Is Insured Unemployment So Low?," *Brookings Papers on Economic Activity*, 1:1983, 225-49.
- Clark, Kim B. and Summers, Lawrence H., "Labor Market Dynamics and Unemployment: A Reconsideration," *Brookings Papers on Economic Activity*, 1:1979, 13-60.
- Summers, Lawrence H., "Why is the Unemployment Rate So Very High near Full Employment?," *Brookings Papers on Economic Activity*, 2:1986, 339-83.

# The Role of Wages in the Inflation Process

By ROBERT J. GORDON\*

The Phillips curve was initially formulated as a relationship between the rate of *wage* change and unemployment, yet what matters for stabilization policy is the rate of inflation, not the rate of wage change. The "wage equation," the traditional centerpiece of the aggregate supply sector of large-scale econometric models, may be redundant, misleading, or irrelevant. If price changes precisely mimic wage changes, then the wage equation is redundant, since all that is needed to guide stabilization policy is a Phillips curve expressed as a relation between inflation and unemployment, with no role for wages. If on the contrary there are systematic differences between the inflation rate and the growth rate of wages adjusted for productivity change, then changes in wage growth may be misleading as an indicator of inflation behavior and wage equations by themselves may yield inaccurate estimates of the natural rate of unemployment. If these systematic differences exist, yet wage changes do not make a statistically significant contribution to the explanation of inflation behavior, then wage equations are irrelevant to the crucial research task of estimating the natural rate of unemployment, the central macroeconomic concept that indicates to policymakers the available scope for economic expansion.

The most striking result in this paper is that wage changes do not contribute statistically to the explanation of inflation, with the profound implication that the aggregate supply process in the United States is characterized by a *dichotomy*: inflation depends on past inflation, not past wage changes. Deviations in the growth of labor cost from

the path of inflation cause changes in labor's income share, and changes in the profit share in the opposite direction, but do not feed back to the inflation rate. There is no support for the age-old structural interpretation, common to almost all Keynesian large-scale econometric models, of wage equations as representing labor market behavior and of price equations as reflecting the "markup" pricing decisions of business firms. The Phillips curve wage equation matters only for the distribution of income, and the markup pricing hypothesis is dead.

The quantitative evidence supporting these surprising assertions takes the form of new estimates of econometric equations for both prices and wages extending over the full 1954–87 period and several subperiods. The format of the specification differs from any of my previous work by allowing lagged wage and price changes to enter symmetrically, rather than excluding one or the other from equations for wage and price change. An interesting transformation of the inflation equation shows that the role of wages reduces to the question as to whether changes in labor's income share mainly affect inflation or the profit share of income.

As a by-product, the new evidence yields numerous additional findings. The U.S. natural unemployment rate is still 6 percent, with no decline in the 1980's in response to beneficial demographic shifts. The U.S. inflation process is stable, with no evidence of structural shifts over the 1954–87 period. But the wage process is not stable: low rates of wage change in 1981–87 cannot be accurately predicted by wage equations estimated through 1980. However, rather than representing a "new regime," wage behavior in the 1980's, and the accompanying decline in labor's income share, can be interpreted simply as reversing the even larger increase in labor's share that occurred between 1965 and 1978.

\*Northwestern University, Evanston IL 60208, and NBER. This research is supported by the National Science Foundation. I am grateful to Dan Shiman for excellent research assistance and to Mark Watson for helpful discussions.

# I. Labor's Share and the Inflation Rate

## A. Specification of the Wage and Price Equations

A general specification of an equation for the rate of price change ( $p_t$ ) is

$$(1) \quad a(L)p_t = b(L)w_t + c(L)X_t + d(L)z_t + e_t,$$

where lowercase letters designate first differences of logarithms, uppercase letters designate logarithms of levels,  $w_t$  is the growth rate of a wage index,  $X_t$  is an index of excess demand (normalized so that  $X_t = 0$  indicates the absence of excess demand),  $z_t$  is a vector of other relevant variables, and  $e_t$  is a serially uncorrelated error term. The vector  $z_t$  includes "supply shift" or "supply shock" variables that can alter the rate of inflation at a given level of excess demand (for example, changes in the relative price of energy), and all components of  $z_t$  are expressed as first differences and normalized so that a zero value of any element of  $z_t$  indicates an absence of upward or downward pressure on the inflation rate. Equation (1) is a general form that can encompass equations in non-structural VAR models or, with restrictions, can be made to resemble traditional "structural" price and wage equations.

The coefficients  $a(L)$ ,  $b(L)$ ,  $c(L)$ , and  $d(L)$  are polynomials in the lag operator  $L$ , and  $a(L)$  is normalized so that its first element equals unity.<sup>1</sup> With this normalization, the term  $a(L)p_t$  can be rewritten as

$$(2a) \quad a(L)p_t = p_t + a'(L)p_{t-1},$$

and, similarly,

$$(2b) \quad b(L)w_t = b_0w_t + b'(L)w_{t-1}.$$

Substituting (2a) and (2b) into (1), we have a somewhat more transparent version of the price equation:

$$(3) \quad p_t = -a'(L)p_{t-1} + b_0w_t + b'(L)w_{t-1} + c(L)X_t + d(L)z_t + e_t.$$

Here we see that the price equation includes not just lagged values of price and wage change, but also the *current* value of wage change.

What about the wage equation? The price equation written in the form of (3) has the startling implication that *there is no such thing as a separate wage equation*. Equation (3) is a price equation and a wage equation at the same time, as can be seen when (3) is renormalized as follows:

$$(4) \quad w_t = -(1/b_0) \times [b'(L)w_{t-1} - p_t - a'(L)p_{t-1} + c(L)X_t + d(L)z_t + e_t].$$

Thus, without further restrictions, the "price equation" (3) and the "wage equation" (4) are alternative "rotations" of the same equation.

Two main approaches are available to identify separate wage and price equations. First, different sets of  $X_t$  and  $z_t$  variables could be assumed to enter the price and wage equations. However, this is implausible a priori, since any variable relevant as a determinant of price change may also be relevant for participants in the wage-setting process, and vice versa for prices. An alternative approach is to restrict the contemporaneous coefficient on  $w_t$  in the price equation or on  $p_t$  in the wage equation, since it is highly likely that there is a contemporaneous correlation between  $w_t$  and the error term  $e_t$  in (3) or between  $p_t$  and  $e_t$  in (4). The contemporaneous coefficient could be restricted to a particular positive fraction, for example, 0.3 as in Blanchard (1986), or to zero in one of the two equations (for example, the wage equation in my previous

<sup>1</sup>Up to this point, the notation and normalization follow Olivier Blanchard (1987), except for the distinction here between demand and supply variables, and except for my assumption that the error term is serially uncorrelated.

papers, say, 1985). In the estimated equations in this paper, the price and wage equations are placed on an equal footing by excluding the contemporaneous wage or price term from both equations, that is,

$$(5) \quad p_t = a^p(L)p_{t-1} + b^p(L)(w - \theta)_{t-1} + c^p(L)X_t + d^p(L)z_t + e_t^p,$$

$$(6) \quad (w - \theta)_t = b^w(L)(w - \theta)_{t-1} + a^w(L)p_{t-1} + c^w(L)X_t + d^w(L)z_t + e_t^w,$$

where an identical set of  $X_t$  and  $z_t$  variables is entered into each. The wage change variables ( $w_t$ ) in (3) and (4) have been replaced in (5) and (6) by wage change minus the change in labor's average product  $(w - \theta)_t$ , that is, the change in unit labor cost.

#### B. Changes in Labor's Share and the Role of the Wage Equation

Hiding inside equation (5) is an interesting relationship between inflation and changes in labor's income share. In the notation of (5) and (6), the change in labor's share ( $\Delta S_t$ ) is defined as

$$(7) \quad \Delta S_t = w_t - \theta_t - p_t.$$

The effects of changes in labor's share in the inflation equation are more transparent if (5) is rewritten in the following form, adding and subtracting the contribution of lagged inflation,  $a^p(L)p_{t-1}$ . Then we have

$$(8) \quad p_t = [a^p(L) + b^p(L)]p_{t-1} + b^p(L)(w - \theta - p)_{t-1} + c(L)X_t + d^p(L)z_t + e_t,$$

which, from (7), implies that lagged changes in labor's share are a determinant of the rate of inflation:

$$(9) \quad p_t = [a^p(L) + b^p(L)]p_{t-1} + b^p(L)\Delta S_{t-1} + c(L)X_t + d^p(L)z_t + e_t.$$

An equation for the change in unit labor

cost, written in parallel form to (8), is

$$(10) \quad (w - \theta)_t = [a^w(L) + b^w(L)](w - \theta)_{t-1} - a^w(L)(w - \theta - p)_{t-1} + c^w(L)X_t + d^w(L)z_t + e_t^w.$$

The effect of a change in labor's share depends on the sum of coefficients ( $\sum b_t^p$ ) in (9). If that sum is zero, then wage changes are irrelevant for inflation, meaning that the counterpart of any increase in labor's income share is a profit squeeze rather than upward pressure on the inflation rate. If that sum is a positive fraction between zero and unity, then an increase in labor's income share becomes another form of supply shock, that is, the  $\Delta S$  and  $z$  terms enter symmetrically. In short, with a positive sum of  $b_t^p$  coefficients, a change in labor's share becomes a source of "cost push" that is on an equal footing with any other type of adverse supply shock, for example, an increase in the relative price of energy or any other variable that causes a positive realization of the  $z_t$  vector. However, if the sum of the  $b_t^p$  coefficients is insignificantly different from zero, this would imply a *dichotomy* between the time-series processes determining the inflation rate and labor's share. Wage behavior would be irrelevant in determining the inflation rate and the natural rate of unemployment, and the wage equation would be of interest only for its description of changes in the distribution of income.

#### C. Interpretations of the Natural Rate

The main focus in this paper is on estimates of equation (8) for price change and (10) for wage change. As a by-product of our main interest in the coefficients on labor's share in the price equation ( $b_t^p$ ), we can use (8) to assess alternative time-series on the natural rate of unemployment. If (8) simply contained lagged inflation and excess demand terms, we could adopt the traditional definition of the natural rate of unemployment ( $U_t^*$ ) as the rate consistent with steady inflation. However, the additional terms in (8) suggest an augmented definition: the natural unemployment rate is consistent with



steady inflation when the net effect of changes in labor's share and of supply shocks is zero.

Due to space limitations, my estimates of (8) and (10) employ a single proxy for the excess demand variable  $X_t$ , the previously developed time-series on the "log output ratio." This is the ratio of actual real GNP to a piecewise log linear trend defined to equal real GNP in selected quarters when the actual unemployment rate is equal to the natural unemployment rate. In turn, the natural unemployment rate series is taken from my paper (1982); it displays a gradual increase from 5.1 percent in the mid-1950's to 6.0 percent in the late 1970's, and is assumed to remain constant at 6.0 percent during 1981-87. Below, I use forecasting tests for 1981-87 based on equations estimated through 1980 to determine whether this natural rate series, and its "dual," the log output ratio, understate or overstate the degree of "slack" in the economy in the 1980s.

## II. The Data and Specification

Equations (8) and (10) are estimated for the period 1954:2 through 1987:3 in the same format as my most recent published results for U.S. quarterly data (1985), except for the free entry of both lagged price and lagged labor cost changes into the price and wage equations in place of my previous practice of excluding one or the other lagged series. As in (8) and (10), where lowercase letters designate rates of change, all variables are entered as first differences in logs, except for the proxy for  $X_t$ , the log output ratio. Data on the fixed-weight GNP deflator is used for the price-change terms, and an index of average hourly earnings adjusted for overtime, employment mix, and fringe benefits is used for the wage-change terms. The productivity growth rate entered into the equations in place of the  $\theta_t$  term is a piecewise linear trend between benchmark years ( $\theta_t^*$ ), and the vector of supply shocks ( $z_t$ ) is entered as in my earlier paper (1985). Further details are provided in the notes to Table 1.

The practical importance of the fringe benefit adjustment and of changes in labor's

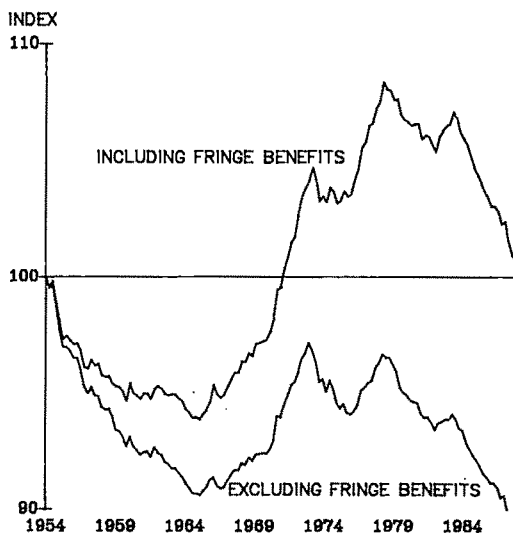


FIGURE 1. INDEXES OF LABOR'S SHARE  
(1954:1 = 100)

share is dramatized in Figure 1. Two indexes of labor's share are shown, calculated simply by cumulating the difference  $(w - \theta^* - p)$ , and expressing the cumulated index on the basis 1954:1 = 100. Of the two share indexes, that appearing as the lower index in Figure 1 is based on average hourly earnings before adjustment for fringe benefits, and the upper index includes the fringe-benefit adjustment. Thus the upper index is based on exactly the same data as the regression equations.<sup>2</sup> The fringe benefit adjustment cumulates to 12 percentage points over the sample period. The fringe-adjusted share index, after declining by 6 percentage points between 1954 and 1965, exhibits a sharp increase of fully 14 percentage points between 1965 and 1978, followed by a 7 point decline during 1978-87 almost back to the starting point. For the full period 1965-87, these up-and-down movements in labor's share occur at an absolute annual rate of 1 percent, large enough for estimated wage-change equations to be-

<sup>2</sup> These indexes do not yield precisely the same index of labor's share as could be obtained directly from the national income and product accounts, because 1) my calculation is based on trend rather than actual productivity, and 2) my wage index refers to the nonagricultural private economy while my price index refers to the total economy.

TABLE 1—BASIC EQUATIONS FOR QUARTERLY CHANGE  
IN FIXED WEIGHT DEFLATOR AND TREND UNIT  
LABOR COST, UNRESTRICTED VERSION,  
1954:2–1987:3

	Fixed Weight Deflator (1)	Trend Unit Labor Cost (2)
<b>Independent variable</b>		
Fixed-Weight Deflator <sup>c</sup>	0.99 <sup>b</sup> (8.0)	—
Trend Unit Labor Cost <sup>c</sup>	—	1.06 <sup>b</sup> (4.8)
Labor Cost/Deflator <sup>c</sup>	0.47 (16.6)	−0.22 (4.4)
Output Ratio	0.17 <sup>b</sup>	0.21 <sup>b</sup>
Productivity Deviation	−0.19 <sup>a</sup>	−0.03
Food and Energy Price Effect	0.33	0.23
Relative Import Price	0.06	0.07 <sup>a</sup>
Relative Change in Consumer Prices	0.08	−0.02
Effective Minimum Wage	0.03	−0.00
Effective Payroll Tax	0.19	−0.18 <sup>d</sup>
Effective Personal Tax	0.06	0.18
Effective Indirect Tax	0.51	0.21
Nixon Controls "On"	−0.84	0.17
Nixon Controls "Off"	1.19	0.21
Summary statistic		
$\bar{R}^2$	0.854	0.913
Sum of Squared Residuals	75.0	53.2
Standard Error	0.963	0.811

Notes: The dependent variable in column 1 is the quarterly change in the fixed-weight GNP deflator. The dependent variable in column 2 is the quarterly change in "trend unit labor cost," defined as the quarterly change in the fringe-adjusted BLS average hourly earnings index for the private economy minus the quarterly change in a productivity trend, defined as a piecewise linear trend of the level of nonfarm private business output per hour between the benchmark quarters of 1954:2, 1964:3, 1972:1, 1978:4, and 1986:4. The fringe adjustment consists of multiplying the BLS average hourly earnings index by the ratio in the National Income and Product Accounts of total compensation to total wages and salaries. All rate-of-change variables are expressed as annual rates, that is, as the quarterly change in the natural log times 400.

The coefficients shown on the first three lines are sums of coefficients on six sets of lagged variables. The first is the average of lags 1–4, the second is the average of lags 5–8, and so on through the sixth variable, the average value of lags 21–24.

Designating "0" as the current quarter, lag lengths for the other variables are chosen as follows: 0–4: Output ratio, food-energy effect, all tax variables; 0–1: Productivity deviation; 1–4: All others.

The Nixon controls "on" dummy variable, taken from my paper with Stephen King (1982) and my paper (1985), is entered as 0.8 for the five quarters 1971:1–1972:3. The "off" variable is equal to 0.4 in 1974:2 and 1975:1, and to 1.6 in 1974:3 and 1974:4. The respective dummy variables sum to 4.0 rather than 1.0 because the dependent variable in each equation is a quarterly change expressed as an annual rate.

<sup>a</sup>Significance of sums of coefficients at the 5 percent level.

<sup>b</sup>Significance (as in a) at the 1 percent level.

<sup>c</sup>Number in parentheses is mean lag.

<sup>d</sup>See text, Sec. III.

have quite differently, and to imply a different natural rate of unemployment, than estimated price-change equations.

### III. Regression Results

Table 1 presents the basic regression results for the price and wage equations cor-

responding to (8) and (10), where the log output ratio is used as the excess demand variable. In keeping with the view that any relevant variable could in principle influence price or wage behavior, I include in both the price and wage equations *all* of the supply shift variables ( $z_t$ ). In the complete price equation (col. 1), the sum of coefficients on lagged inflation is almost exactly unity, indicating that the theoretical presumption of unity can be accepted. An equally important, and perhaps more surprising result, is that the sum of coefficients on the lagged labor's share variable ( $w - \theta - p$ ) is insignificantly different from zero, with a 0.12 significance value on the sum of coefficients and a 0.24 value on an exclusion test of this variable. In parallel fashion, the labor's share variable in the labor cost equation in column 1 is also insignificant, with a 0.32 significance value on an exclusion test. These results, then, support the "dichotomy hypothesis" that wages do not matter for price behavior and vice versa.

Looking now at the other variables, the sum of coefficients on the output ratio terms is highly significant in both columns. The magnitude of these sums of coefficients is lower than in my equivalent past research, a change which stems entirely from data revisions in the national accounts. Of the supply shifts, the sums of coefficients that are significant are those for the deviation of productivity growth from trend in the price equation and the relative import price in the labor cost equation. The payroll tax in the labor cost equations is highly significant, note, however, that it enters in the form of a positive coefficient followed by a string of negative coefficients, yielding an insignificant sum. This pattern can be interpreted as suggesting that an increase in the effective payroll tax initially raises labor cost, but that subsequently the tax is "backward shifted" from employers to workers.

#### A. Implications for the Natural Rate of Unemployment

The log output ratio entered into all of the regression equations thus far in the paper is constructed as the "dual" to a hybrid natural unemployment rate series ( $U_t^{G*}$ ) used in

previous research. For readers of this paper, then, the natural rate series "drops from the sky," and an assessment of this series is now overdue. Two techniques are used to provide this assessment. First, equations are re-run with dummy intercept shift terms for 1963-68, 1969-74, 1975-80, and 1981-87, and the coefficients on these shift terms are examined for significant values. A significant positive value would indicate that price and/or labor cost change was faster than the equation can explain, implying an underestimate of the natural unemployment rate, while a significant negative value would imply the opposite. Since my hybrid natural rate series ( $U_t^{G*}$ ) assumes a 6.0 percent natural unemployment rate after 1980, the optimistic view that the natural unemployment rate has fallen from 6.0 to perhaps 5.0 percent in recent years would be supported by a significantly negative coefficient on the intercept shift coefficient for 1981-87.

The top section in Table 2 displays the intercept shift coefficients that are intended to measure shifts in the natural rate from the assumed series. None of these coefficients are significant. In particular, the coefficient in the price equation for the 1981-87 interval is very close to zero. The relatively large (but insignificant) positive coefficient for 1969-74 in the wage equation and corresponding negative coefficient in the price equation is the counterpart of the increase in labor's share in that interval evident in Figure 1. A joint significance test on the four intercept shift coefficients indicates a very low level of significance in the price equation, and a marginal 0.11 level in the labor cost equation.

The lower section of Table 2 provides summary statistics on dynamic simulations for 1981-87 of equations estimated for 1954-80. All simulations are dynamic in the sense that lagged price and labor cost terms are generated endogenously. The three summary statistics are 1) the error in the last 4 quarters of each 27-quarter simulation, providing a measure of the simulation's "drift" in 1986-87; 2) the average error, indicating the overall bias of the simulation, and 3) the simulation's root-mean-squared-error (RMSE), measuring its overall accuracy.

TABLE 2—PERFORMANCE OF LOG OUTPUT RATIO  
MEASURE OF EXCESS DEMAND AS MEASURED BY  
CONSTANT SHIFT TERMS AND BY  
POST-1980 SIMULATION ERRORS

Sample Period	Complete Price Equation (1)	Complete Labor Cost Equation (2)
<b>1954-87:</b>		
Coefficients on Shift Dummies:		
1963:1-1968:4	-0.20	0.31
1969:1-1974:4	-0.56	0.65
1975:1-1980:4	-0.11	0.45
1981:1-1987:3	-0.06	-0.48
Joint Significance of Dummies	0.88	0.11
<b>1954-80:</b>		
Dynamic Simulation Errors:		
Average 4-Quarter Error for 1987:3	-0.93	-1.72
Average Error for 1981:1-87:3	-0.27	-1.76
Root Mean Square Error for 1981:1-87:3	1.15	2.09

The mean error for the price equation is extremely low, only -0.27 percent at an annual rate, indicating only a slight tendency to overpredict the inflation rate. Drift began in 1986-87, leading to a larger -0.97 percent error in the year ending in 1987:3. The RMSE is 1.15 percent, only a bit higher than the sample period standard error (for the 1954-80 interval) of 1.04 percent. Since it is the price equation that matters for the natural rate of unemployment, the low mean error for 1981-87 suggests that my assumed natural rate series remains accurate for this period. There is no evidence that the natural rate has drifted down below 6 percent; this is particularly true when the labor cost variable is excluded from the price equation, in which case too little inflation is predicted in 1986-87. The large overpredictions of inflation in the labor cost equation in Table 2 are the counterpart of the decline in labor's share in the 1980's and have no implications for the natural rate of unemployment.

#### IV. Conclusion

Traditionally wage equations of the Phillips curve variety are the central element explaining inflation in large-scale Keynesian econometric models. Price changes are specified as determined by a "markup" price equation and have little life of their own, mainly mimicking wage changes. Such a view of the inflation process is rejected by this paper. A relatively unrestricted equation for

price change can be converted into a form in which wage changes enter only in the form of lagged changes in labor's share. When the labor's share variable is statistically insignificant, as reported here, *wage behavior becomes irrelevant for inflation*. Differences in the behavior of labor cost and inflation imply changes in labor's income share which alter the profit share of income in the opposite direction.

The paper also concludes that price changes are irrelevant for wage changes, that is, that both prices and labor costs live a life of their own. Here the evidence is less clear than in the price equations; an alternative version that allows the distribution of coefficients on lagged prices and wages to shift after 1967 indicates that either prices or wages provide an adequate explanation of wage changes. None of these equations, however, provide any substantive explanation of the sharp increase in labor's income share during 1965-78 or its subsequent decline. Thus the results are consistent with those who claim that the decade of the 1980s has witnessed a "new regime" in wage formation; virtually all of my estimated wage equations show a marked tendency to overpredict wage change for 1981-87 on the basis of coefficients estimated for 1954-80. That is, from the point of view of the equations, wage changes in 1981-87 have been too low.

No evidence is provided here on the causes of such a new regime in wage behavior in which labor's share has fallen, nor indeed on the causes of the old regime in which labor's share rose from 1965 to 1978. In fact, the new regime may just represent the unwinding of the old regime. It is notable that the timing and extent of this change in labor's share parallels that which occurred in most European countries at the same time, leading to skepticism that factors unique to the United States, for example, foreign competition, deregulation, and waning union power, have caused the turnaround in labor's share. The parallel timing of the U.S. and European rise and fall of labor's income share may also throw cold water on those who have stressed unique aspects of European wage behavior as an underlying cause of high European unemployment in the 1980's.

Given its successful past performance, it is interesting to use this paper's inflation equation to generate predictions for the future. If we make the crucial assumption that a supply-shift variables have future effects netting out to zero, we can run dynamic simulations of the price-change equation starting in 1987:4 for two different assumed paths of the unemployment rate.<sup>3</sup> The first path calls for unemployment to remain at 6.0 percent forever, and the second for unemployment to decline to 5.0 percent by 1988:4 and to remain there forever. The 6 percent unemployment path is consistent with steady inflation forever of 3.5 percent, almost exactly the inflation rate for the four quarters ending in 1987:3. A steady acceleration of inflation is implied by the 5 percent unemployment path, amounting to 1.1 points of extra inflation after five years and 2.4 points after ten years (i.e., the inflation rate reaches 6 percent in 1997).

Some may view this modest acceleration of inflation as a small price to pay for reduction of unemployment by 1 percentage point, which would yield roughly \$100 billion *per year* in extra GNP at today's prices or more than \$1 trillion over the 1987-97 decade. But these proponents of demand stimulus are obliged to indicate when, anyhow, the steady acceleration of inflation is to be stopped. Those who would prefer a path of steady inflation can translate the 6 percent unemployment forecast into a recommendation that the Fed maintain a steady 5.9 percent growth rate of nominal GNP consisting of 3.5 percent inflation plus 2.4 percent for real GNP, the latter being the growth rate of natural real GNP between 1979 and 1987.

<sup>3</sup>The assumed unemployment paths are converted into assumed paths for the log output ratio by using the Okun's Law coefficients from my 1984 paper, which indicate that the long-run response of unemployment to a change in the log output ratio is 0.45.

## REFERENCES

- Blanchard, Olivier J., "Empirical Structural Evidence on Wages, Prices, and Employment," Working Paper 431, MIT, September

ber, 1986.

\_\_\_\_\_, "Aggregate and Individual Price Adjustment," *Brookings Papers on Economic Activity*, 1:1987, 57-109.

Ordon, Robert J., "Inflation, Flexible Exchange Rates, and the Natural Rate of Unemployment," in Martin N. Baily, ed., *Workers, Jobs, and Inflation*, Washington: The Brookings Institution, 1982, 89-151.

\_\_\_\_\_, "Unemployment and the Growth of

Potential Output in the 1980s," *Brookings Papers on Economic Activity*, 1984:2, 537-64.

\_\_\_\_\_, "Understanding Inflation in the 1980s," *Brookings Papers on Economic Activity*, 1985:1, 263-99.

\_\_\_\_\_, and King, Stephen R., "The Output Cost of Disinflation in Traditional and Vector Autoregressive Models," *Brookings Papers on Economic Activity*, 1982:1, 205-42.

# THE MEASURE AND CHARACTER OF AMERICAN UNEMPLOYMENT<sup>†</sup>

## The Measurement of Unemployment

By JANET L. NORWOOD\*

For more than forty years, the U.S. government has produced a monthly estimate of unemployment. The measurement comes from the *Current Population Survey (CPS)*, one of the most comprehensive of all the household surveys in this country, and generally recognized as one of the best labor force surveys in the world. Each month, the estimate is widely discussed. A change in the unemployment rate can trigger a change of direction in national economic policy, or a flurry of activity in the financial markets. Policymakers interpret the numbers with great care, and the general public await their release so that they can evaluate policy. The unemployment rate usually makes the nightly TV news and the front page of the morning newspaper.

In spite of this intense interest, the official definition of unemployment is not always fully understood. Many people think the statistic comes from a count of those receiving unemployment compensation; others are surprised that no one in the survey is specifically asked whether or not he or she is unemployed. The definition is based on activity—job search. Those persons in the survey who have not worked during the survey week, are available for work, and have looked for work during the preceding 4 weeks are counted as unemployed. Unless the respondent fits this activity definition, he or she does not meet the official definition of unemployment.

The official definition of unemployment has long been a source of debate among

labor market analysts. Twenty-five years ago the Gordon Committee (President's Committee, 1962), appointed by the president to review unemployment statistics, recognized the limitations of the unemployment measure. The Gordon Committee recommended modifications that were subsequently adopted (1967) to reduce ambiguities in the definition. They also recommended further research to produce measures that would be easier to understand. Although many have complained about the definition, however, few have attempted to provide a conceptual framework for measurement.

Economists have done research on the interaction of aggregate demand and the labor market and on the relationship between the unemployment rate and business cycle developments (see Thomas Kniesner and Arthur Goldsmith, 1987). The labor market literature includes work on models aimed at explaining labor force entry, unemployment and real wages, job search, structural and frictional unemployment, as well as looking at the effects of unemployment on future labor market experience, economic hardship and other social issues relating to unemployment.

In spite of this body of concepts found in the economic research, however, little attention has been focused on establishment of a framework for the measurement of unemployment. This problem generally is left to those responsible for survey design, who must deal with all users of the data system—those who criticize the definition of unemployment as well as those who approve of it.

Criticism of the definition tends to focus more on the uses of the data—and the problem of focusing on that use—than on the underlying theoretical framework for the definition. It is said that the unemployment

<sup>†</sup>*Discussants:* Daniel Hamermesh, Michigan State University; Donald Nichols, University of Wisconsin-Madison; Robert M. Solow, MIT.

\*Commissioner of Labor Statistics, U.S. Department of Labor, Washington, D.C. 20212.

measure is essentially a reflection of potential labor supply, but (as Clarence Long, 1947, emphasized long ago), the potential labor force could include many not counted as employed or unemployed. At one end of the spectrum are those who are actively looking for work. At the other end are those who are and will remain entirely out of the labor force. Between these two extremes are those currently out of the labor force who might want to enter it if certain conditions or circumstances occurred. The line that divides active job seekers from those not in the labor force at all is really difficult to draw with precision and finality. The Levitan Commission (National Commission, 1979), which reviewed the unemployment data system in 1979, spent considerable time discussing this issue. Although not recommending a change in the official definition of unemployment, the Commission recognized that the data system did not provide information on the potential for entry into the paid labor force. Indeed, information on marginal attachment to the labor force is not available, nor are data on reservation wage collected.

Others whose major concern is the measurement of the welfare of the population find the unemployment definition flawed because it does not effectively measure labor market related hardship (Sar Levitan and Robert Taggart, 1974). Job loss can result in economic, social, and psychological hardship, but the unemployment data in themselves do not measure economic distress. Although unemployment is frequently used as a proxy for a welfare measure, it is not really satisfactory for that purpose. In fact, in some situations, the working poor may suffer greater economic hardship than the unemployed, who in some cases are eligible for unemployment and other financial benefits. In addition, the current data system provides insufficient help to those wishing to study the psychological, social, or predictive effects of unemployment.

The economics profession has made much use of the unemployment rate and other labor force information produced from the CPS. Although their research has used a number of theoretical concepts to explain observed relationships in the labor market,

economists generally have shown little interest in the problems of measurement, nor have they thought much about the development of a comprehensive conceptual framework for the design of the labor force survey.

While we at the Bureau of Labor Statistics hope that economists in the future focus more of their attention on this challenge, we must continue to provide data that we believe are required to understand what is happening in the labor market. Recently, we began a long-range planning effort to determine what improvements should be made in the labor force survey to produce the kind of data needed to understand labor market developments of the future.

We believe that the redesign of the CPS after data from the 1990 Census become available should be comprehensive and carefully planned. It should include innovative approaches to sample design and estimation design, use of modern technology in the data collection process, and the application of cognitive research in questionnaire design.

In addition, however, we want to ensure that the data produced are not simply improved but expanded and changed in ways that will make them more relevant to the needs of future users. The phenomena measured by our labor force survey are constantly changing. Survey output must also change if we are to provide a data base of labor market information to cope with the problems of the future. This paper will briefly discuss three areas where improvement is needed—designing questionnaires, expanding individual state data, and improving expanded longitudinal data.

### I. Cognitive Research on Questionnaire Design

In the past, far too little attention has been given to the possible errors caused by a misunderstanding by the respondent of what we are trying to collect. Survey researchers have learned that the words used in drafting a question are important, that minor changes in wording and small changes in the placement of questions can affect survey results. For example, last year, we made a small change in the wording of a CPS question—

the question on availability for work was asked in a more direct manner. We found (and publicly announced) that the small change (even after previous careful testing had shown no effect from the recasting of the question) resulted in a tenth of a percentage point understatement of the unemployment rate for the month (January).

Professionals at the BLS and the Census Bureau have jointly reviewed the cognitive issues in the *CPS* questionnaire and have prepared a research agenda for the future. This would involve testing new approaches to those parts of the questionnaire where wording may appear ambiguous or where the task of recall may be especially complex for the respondent. An example identified by the task force is the very first question in the *CPS* employment series. The respondent asked what he or she was "...doing most of last week—working, keeping house, going to school or something else?" must interpret the question before giving a response. Does the question ask a) which activity took up most of last week's time, or b) what activity was the one at which most of last week's time was spent? Probably no single activity took up most of the week (i.e., 168 hours), and sleeping probably took more time than anything else. But neither is intended; the question is intended to find out about the respondent's status or behavior.

Cognitive research in survey design will bring together researchers from different disciplines to improve the process which thus far has been left primarily to the economists and statisticians. We hope to establish a laboratory for testing questionnaire wording and to contract research by social psychologists, linguists, and sociologists to investigate the interaction that occurs between interviewer and respondent, the interpretation of words and thoughts encountered by respondents, as well as to understand better the cognitive processes of comprehension and recall.

## II. Expansion of Data for Individual States

Unemployment has always differed from one area of the country to another—in periods of expansion as well as in periods of

recession. But in recent years, we have become more aware of these differences. In part, this is because U.S. industry, which tends to be concentrated in particular geographic areas, is undergoing considerable change. As plants have been shut down or consolidated, unemployment has shot upward. All industries have not been equally affected, however, and the labor market distress suffered has often differed from one place to another.

Estimation of unemployment data for states and areas of the country in a timely and accurate manner is difficult at best. In most cases, the data are developed from a combination of survey and administrative data and cannot be produced as quickly and accurately as the data for the national employment situation.

We know, however, that the country's overall unemployment rate is very much affected by changes in industry mix. Our ability to analyze the labor market data each month is very much limited by the fact that data for only 11 states are accurate enough for monthly publication directly from the *CPS*. Unemployment data for the remaining smaller states are derived from a number of independent data sources as well as the *CPS* and are not available until a month after the national data have been released.

Although BLS-Census Bureau joint planning in this area is still at a very early stage, we are considering the possibility of expanding state data in the next redesign of the *CPS*. If these plans work out—and if OMB and congressional approval are secured—the survey could be expanded from its current size of approximately 60,000 households to approximately 100,000 in order to permit calculation of reasonably reliable estimates for each of the 50 states.

## III. Improved Longitudinal Data

The *CPS* has been processed as though it were a new survey every month. Because the purpose of the *CPS* has been to provide a snapshot of the labor market each month, little effort has been made to place the information collected in the current month in a longitudinal framework to determine its con-



sistency and accuracy. It is clear, however, that the interpretation of monthly change and our understanding of labor market behavior in general could be enhanced if we were able to track the flow from one labor market status to another.

In fact, the *CPS* does have some aspects of a longitudinal survey; unfortunately, they have not been given adequate attention. *CPS* respondents can be in the survey for as long as 16 months. The households in the sample are interviewed for 4 consecutive months, dropped out of the survey for 8 months, and again are interviewed for another 4 months. Thus, the capability to document the transitions between labor market status—employment, unemployment, and out of the labor market—are to a considerable extent available in the current *CPS* design.

But we should not underestimate the task of improving the longitudinal aspects of the *CPS*. Although development of improved longitudinal data will be one of our major priorities for the *CPS* of the future, we at BLS recognize that the task will require a complete redesign of the computer processing system as well as improved coding and editing procedures. In addition, procedures for ensuring consistency of response need to be developed and tested. The statisticians working on the *CPS* have quite rightly been concerned about the possible bias that might occur in the estimates if the interviewer had access to information about the respondent from earlier interviews.

These efforts need to be directed toward improving two areas of the data, each of which is important for analysis. The first involves the use of aggregate data—the gross flows—and the other involves use of the micro data—matching responses for the same individual over time. The gross flows data have been produced for many years, but because they are at times inconsistent with the cross-sectional data produced each month, they have not been very useful. In 1984, the BLS and Census Bureau sponsored a conference on statistical issues involved in improving the gross flows material (see the *Proceedings...*, 1985). Although we recognize the complexity of some of the technical issues involved, the BLS believes that some

progress can be made in this area in the future.

Considerable work has also been done in development of tapes containing micro-matched responses, but it is clear a very real effort needs to be made to solve some of the problems which prevent full use of those data. Problems exist in the processing of the data, and response variability and conditioning problems exist which we do not fully understand. In addition, we need to learn more about the effect of movers on the *CPS* estimates. The *CPS* is based on a sample of household addresses; when a *CPS* respondent moves out of the address in the sample, he is not followed for the survey. Instead, the person moving in becomes a new respondent in the survey. One way to learn more about this problem, and to secure more information at the same time, would be to ask the *CPS* questions to a small (perhaps 3,000 households) additional sample of respondents over a period of 2.5 to 3 years. We are considering this possibility as one of many projects to improve the *CPS* in the 1990's. Such a longitudinal companion to the *CPS* might be difficult to implement (because of possible sample attrition) and could be costly. But it would provide a very rich body of data to help to understand labor market change.

#### IV. Conclusion

The *Current Population Survey* provides a very rich body of demographic labor market information; in many ways, the *CPS* provides a management information system for economic policy. The redesign of the 1990's will provide both a very great challenge and a marvelous opportunity. But if we are to make the kinds of changes that I believe are needed for us to understand the labor market of the future, the groundwork must be put in place very soon. The use of new technology in collection and processing provides an unusual opportunity for improvement of the survey output. A great deal of economic research in the past has been based on *CPS* data. As we develop our plans for the 1990's redesign, the Bureau of Labor Statistics would welcome comments on future direc-

tions for the CPS from the economics and statistics professions.

#### REFERENCES

- Kniesner, Thomas J. and Goldsmith, Arthur H., "A Survey of Alternative Models of the U.S. Labor Market," *Journal of Economic Literature*, September 1987, 25, 1241-80.
- Levitan, Sar A. and Taggart, Robert, *Employment and Earnings Inadequacy: A New Social Indicator*, Baltimore: Johns Hopkins University Press, 1974.
- Long, Clarence D., "The Concept of Unemployment," *Quarterly Journal of Economics*, November 1942, 57, 27-30.
- National Commission on Employment and Unemployment Statistics, *Counting the Labor Force*, Washington: USGPO, 1979.
- President's Committee to Appraise Employment and Unemployment Statistics, *Measuring Employment and Unemployment*, Washington: USGPO, 1962.
- U.S. Department of Commerce, Bureau of the Census, and U.S. Department of Labor, Bureau of Labor Statistics, *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, Washington: USGPO, 1985.

# What Is So Natural About High Unemployment?

By RICHARD S. KRASHEVSKI\*

The prevalent but incorrect belief that the United States now has a natural unemployment rate around 6 percent fosters an unwillingness among the nation's leaders to promote economic growth sufficient to cut joblessness below that rate. Earlier, at least through the 1960's, a consensus held that economic growth could bring the nation's unemployment rate down to 4 percent, and even lower when combined with policies to address structural problems. In the last half of the 1960's, unemployment remained below 4.0 percent for four consecutive years—one of the best sustained performances in the post-World War II period. But in the 1970's and 1980's, joblessness climbed to higher levels at each peak and trough and, even though the 1978 Humphrey-Hawkins Act required the achievement of 4 percent unemployment, the consensus to do so fell apart.

Several reasons have been advanced to explain the deterioration and justify an upward redefinition of the minimum unemployment rate achievable through policies that raise aggregate demand. If examined carefully, however, the arguments just don't stand up. Among such ideas are the following: 1) women and youth entered the labor force in great numbers, and allegedly have naturally higher rates of unemployment; 2) joblessness cuts, in particular below the so-called normal rate, supposedly generate wage and price increases which cause recessions and rising joblessness; and 3) social programs and labor legislation ostensibly reduce beneficiaries' needs and desires to work and also make firms hesitant to add workers.

The proposed reasons at most account for a minor part of the climb in joblessness, and

do not prevent the attainment of 4 percent unemployment through policies that promote strong economic growth. Nonetheless, structural unemployment is a serious problem, and its alleviation calls for industrial and trade policies, job and training programs, antidiscrimination enforcement, and other special initiatives. Such efforts will add momentum as economic growth propels the nation towards full employment.

## I. Demographic Changes

Large increases in the work force participation of youths born during the post-World War II baby boom, along with the enormous number of women entering the job market, are believed to have pushed up the overall unemployment rate for a given state of the economy.

To evaluate the relevance of demographic explanations for the rise in joblessness, an annual series of constant-weight unemployment rates was calculated. The measure is similar to those that various authors have employed; George Perry (1970) was one of the first to explore such unemployment rate adjustments, and Paul Flaim (1979) provides a useful overview of the approach.

In the measure used here, the labor force is separated into twelve categories based on age, race, and sex. Each group's actual unemployment rate in a given period is multiplied by its average labor force share of the 1960's. The twelve resulting products, one for each group, are then summed. The end result is a series of yearly constant-weight unemployment rates from which the effects of changes in the demographic composition of the labor force are purged. The difference between the actual unemployment rate and the constant-share rate is ascribed to the shift in labor force composition.

Young people and women, through changes in their labor force shares and through their unemployment rates compared

\*Department of Economic Research, AFL-CIO, Washington, D.C. 20006. I thank Rudolph A. Oswald for important suggestions.

to other groups, have much to do with the demographic gap's evolution and size. When, for example, teenagers' labor force share rose from its 7.8 percent 1960's average to a 9.6 percent peak in 1974, the gap increased and reached its high the following year. The labor force presence of teenagers then steadily receded until, at 6.7 percent in 1986, it had dropped all the way back to the 1950 level. Similarly, the demographic gap declined to a mere 0.1 percent by 1986.

The growth in women's labor force participation was even greater, and it is important to note that the increase is not just a recent or transitory phenomenon. The share of females age 20 and older rose 4.7 percentage points from 1947 to 1957, 3.2 points from 1957 to 1967, 3.7 points from 1967 to 1977, and 4.8 points in the last ten years. What should be considered natural is not a particular female labor force share, but the sustained increase in that share.

In the last few years, the rates of joblessness for adult males and females have typically been about the same because of the erosion of manufacturing industries, where males have predominated, and the concentration of women in less cyclical occupations, albeit generally lower paid. As a result, the continued rise in adult women's labor force share tended not to raise, but to reduce the demographic gap in the 1980's.

With the share of youth declining and the margin between adult male and female joblessness diminishing, the effect of changes in the work force from the 1960's has tapered off. By 1986, the gap implied that such demographic differences could make unemployment at most 0.1 percent higher than in the 1960's, during the same phase of the business cycle.

Now, however, unemployment is concentrated especially among workers who lost jobs due to the immense wave of imports in the 1980's. Joblessness in the trade sensitive manufacturing sector was 7.1 percent in 1986 compared to 5.6 percent in 1979. The condition stems partly from an improper macroeconomic policy mix, and may be reduced in part by a reversal of that mix. More focused initiatives are also needed. The problem is not purely a cyclical one, but cannot be

viewed as natural or permanent either. That the locus of more-enduring unemployment difficulties shifts over time illustrates that such problems don't stem from the intrinsic qualities of certain groups. Nonetheless, the problems cannot be ignored.

Other labor force changes, not reflected in the fixed-weight measure used here, should assist unemployment's reduction. One is the rise in education, which enhances productivity and facilitates the absorption of new workers into the labor force. In 1960, 61 percent of 25 to 29-year-olds had at least completed high school. By 1985 that share was 86 percent. Another is the incidence of involuntary part-time work, which, though lower than during the last recession, is currently much higher than in the past. In September 1987, involuntary part-time workers were 5.3 percent of all employees, compared to 3.4 percent in the 1960's, indicating that there is far more slack in labor markets than supporters of a 6 percent joblessness target would acknowledge.

Table 1 provides another perspective on the demographic approach's deficiency in explaining unemployment's sustained deterioration. Over the years since the 1960's, joblessness rose to a new high at each business cycle peak, but demographics explained a progressively smaller part of the rise. Early in the 1970's, demographic shifts seemed to provide a complete accounting for higher unemployment. At the decade's first peak in 1973, both the demographic gap and the rise in unemployment's cyclical minimum equalled 0.4 percentage points.

Such comovements soon broke down, as column 6 of Table 1 reveals. The portion of the unemployment rate's climb attributable to labor force changes dropped, without interruption, throughout the 1970's and 1980's.

Nor can the approach be rescued by using the labor force makeup of the 1950's for the fixed weights. Though each period's gap is greater when 1950's weights are used, the percentage of unemployment's increase explained by the gap suffers the same relentless decline. After reaching a 54 percent maximum in 1973, the explained portion of unemployment's deterioration declined without pause, and labor force changes since the

TABLE 1—ACTUAL AND FIXED-WEIGHT (1960's)  
UNEMPLOYMENT RATES: YEARS OF LOW  
UNEMPLOYMENT DUE TO CYCLICAL PEAKS

(1)	(2)	(3)	(4)	(5)	(6)
1960	5.5	—	—	—	—
1969	3.5	—	—	—	—
1973	4.9	4.5	0.4	0.4	100
1979	5.9	5.4	0.5	1.4	36
1980	7.2	6.7	0.5	2.7	19
1981	7.6	7.2	0.4	3.1	13
1987 <sup>a</sup>	6.3	6.3	0.0	1.8	0

Source: Basic data from U.S. Dept. of Labor, B.L.S. 1/ Fixed weight rate based on 12 age, race, & sex groups. 2/ Unemployment at 1060's cycle peaks = avg 1960 & 1969 rates. 3/ 1987.

Notes: Col. 1 denotes unemployment lows; col. 2 is actual unemployment in percent; col. 2 is fixed-weight unemployment in percent, rate is based on 12 age, race, and sex groups; col. 4 is gap = col. 2 - col. 3; col. 5 is unemployment rise from 1960's 4.5 percent average, and 1969 rates; col. 6 is gap as percent of unemployment rise.

<sup>a</sup>Through 3rd quarter; fixed-weight rate has 4 groups.

1950's now account for just 15 percent of unemployment's change at cyclical peaks.

The composition of the work force is always changing, as is each group's unemployment rate compared to others'. Such constant transformation, and demographics' inability to explain high joblessness, makes it difficult to believe that any one composition is natural, or that a natural rate based on demographics even exists.

## II. Unemployment and Inflation— A Weak Link Crumbles

The view that a 6 to 7 percent range is the lowest that unemployment can descend without causing an inflation speedup incorrectly characterizes labor markets as the principal cause of inflation and wrongly extrapolates the conditions of the 1970's to today's economy. Supply shocks, not excess demand, labor market tightness or wage acceleration, set off the major inflationary surges of the 1970's.

In 1973, food shortages and the oil cartel's enormous oil price hike launched the decade's first wave of double-digit inflation. Another huge oil cartel price increase in 1979 initiated a second round of inflation above 10 percent. Workers' purchasing power eroded as wage increases failed to keep up with increases in the general price level. To-

day, a decade and a half after the whole inflationary process began, real average hourly earnings of nonagricultural production and nonsupervisory workers are 10 percent below their 1973 value. Productivity in the nonfarm business sector advanced about 12 percent from 1973 to 1987's third quarter, implying a reduction in the share of income going to production workers.

Recent experience provides compelling evidence that inflation doesn't necessarily rise when joblessness declines. The worst recession since the 1930's pushed the unemployment rate above 10 percent in late 1982 and the first half of 1983. With the economy in recovery, joblessness at first declined rapidly, but then see-sawed in the 7.0 to 7.5 percent range for over two years from May 1984 to September 1986 before moving down sharply to 6 percent by mid-year 1987.

During unemployment's decline, inflation's pace stayed moderate and remarkably stable, contradicting the presumed negative relationship between the two. The consumer price index, for example, posted annual increases (December to December, seasonally unadjusted) of 3.9 percent in 1982, 3.8 percent in 1983, 4.0 percent in 1984, 3.8 percent in 1985, and—mainly from the oil price collapse—an unusually slow 1.1 percent in 1986. Rebounding oil prices temporarily boosted inflation's pace to 4.7 percent through 1987's third quarter, but the rate decelerated in each period (5.3, 4.9, and 4.0 percent, at seasonally adjusted annual rates). While inflation fundamentally stayed in a narrow range, joblessness dropped from nearly 11 to 6 percent.

Particularly in the past few years, the links between unemployment, wages and inflation have been very weak. The present circumstances—an abundance of unused capacity, joblessness far above the feasible minimum, and moderate inflation—all signal that unemployment can be cut further without quickening inflation's pace.

## III. Social Programs and Labor Legislation

Government-sponsored income-support payments and labor legislation are favorite scapegoats for unemployment's climb to

successively higher levels. The most thoroughly maligned programs are unemployment insurance and the minimum wage, even though they were put in place nearly four decades before unemployment began its extended rise.

Unemployment insurance, and other programs that require recipients to claim that they are seeking work, supposedly inflate the official unemployment rate. Attackers argue that many individuals officially enumerated as unemployed are not truly interested in finding work. If income supports were cut, many would take the jobs available to them.

The reality is that unemployment insurance payments are now made to only one-third of the jobless, eligibility requirements are more stringent, and benefits are sharply lower in real value. The ratio of average UI benefits to average weekly earnings is virtually unchanged from thirty years ago. Other social programs, including food-stamps and AFDC, have also been accused of raising unemployment. In the 1980's, however, the programs' real values dropped one-third.

The minimum wage is yet another scapegoat for high unemployment. The wage floor has sharply declined since its last adjustment to \$3.35 in 1981. From 1981 to 1987, the real value of the minimum wage dropped more than one-fourth, and the ratio of the minimum wage to average hourly earnings diminished from 48 to 37 percent. Albert Rees (1986) points out that, if the minimum wage really were a significant determinant of youth unemployment, the drop in the minimum's value relative to other wages should have caused a decrease in the ratio of youth to adult unemployment, but no such decrease took place.

The idea that various social programs undermine the incentive to work or reduce employer hiring is wrong. While relentless budget reductions and the failure to offset price increases left gaping holes in the nation's social safety net, unemployment remained far above earlier norms.

#### **IV. How to Reduce Unemployment to 4 Percent**

The proposition that labor force changes, inflation, and the social programs caused

unemployment's long steep climb, or now preclude its reduction to 4 percent, is incorrect. The 1980's have seen male and female unemployment rates converge, the labor force share of young people decline, inflation moderate, a sharp decline in the real value of the minimum wage, and severe cuts in income-support programs. The only sustainable conclusion is that 6 percent unemployment in 1987 reflects the same intolerable labor market slack that it did two decades ago. Today, as in the 1960's, full employment is an unemployment rate no higher than 4 percent.

To attain 4 percent unemployment, the United States requires a set of policies designed to stimulate demand and reduce the slack that now pervades the economy. With over seven million people officially counted as unemployed, capacity use around 80 percent, and inflation at low and steady rates, stimulative policies are required. The Federal Reserve Board should worry less about inflation's moderate pace, and emphasize economic growth and cutting unemployment.

A shift to a better mix of fiscal and monetary policies, with a smaller budget deficit and a more stimulative monetary stance, should lower real interest rates and bring much better balance to the nation's economy. When reducing the federal deficit, the nation would be best served by limiting the cuts in essential and already weakened domestic programs and, instead, raising revenues through tax changes that bring greater progressivity to the income tax structure.

Even though the huge federal budget deficit has already boosted aggregate demand, far too much spending has been on imports instead of domestic output. The trade deficit lost over two million better-paying more-productive manufacturing jobs. Demand must be redirected from foreign imports to domestic products with trade and exchange rate policies.

The rise in the dollar's foreign exchange value coincided with the initial deterioration in the nation's trade balance. But the drop in the dollar, which began about three years ago in March 1985, hasn't yet significantly improved the trade balance. Instead the trade deficit has deteriorated further, demonstrating a need for stronger measures to redirect

spending to products and services of United States origin.

Two such measures are the adoption of an industrial policy and the strengthening of the nation's trade laws. Industrial policy can bring together representatives of society's different economic sectors to reach a consensus on both aggregate and structural economic goals. Among the many needed trade law improvements are quicker and more forceful action against imports that are either dumped or subsidized, less delay in establishing emergency relief when imports injure industries or workers, and the termination of tax and other incentives for business to move production overseas. Better financial aid, and job search and relocation assistance must be provided to workers who have lost their jobs because of trade.

To assist the move toward full employment, and address the labor market's structural problems, more and better-funded employment and training programs are es-

sential. The cutbacks in such programs, especially in recent years, make reducing unemployment more difficult, and must be reversed.

The United States must restore the achievement of full employment as the foremost goal of national economic policy. Economic conditions permit and social justice demands a national commitment to reduce unemployment to 4 percent.

#### REFERENCES

- Flaim, Paul, "The Effect of Demographic Changes on the Nation's Unemployment Rate," *Monthly Labor Review*, March 1979, 13-23.
- Perry, George, L., "Changing Labor Markets and Inflation," *Brookings Papers on Economic Activity*, 3:1970, 287-308.
- Rees, Albert, "An Essay on Youth Joblessness," *Journal of Economic Literature*, June 1986, 24, 613-28.

# Evaluating the European View that the United States Has No Unemployment Problem

By RICHARD B. FREEMAN\*

A session on unemployment in America? Ridicule! The U.S. has produced 20 million jobs since 1975. If only Europe had America's flexible labor market and "unemployment."  
[Archetypal European economist, circa 1987]

Significant differences between the unemployment and employment experiences of the United States and OECD-Europe have made views like the above popular overseas and led many European observers to look longingly at the American labor market as a paragon of decentralized wage and employment flexibility.

Do the labor market performances of the United States and OECD-Europe support this view? How much of the difference between American and European employment and unemployment can be attributed to differences in labor market "flexibility"?

In this paper I examine these questions. I review the labor market outcomes that have led many Europeans to see the American economy as having no "real" unemployment problem; evaluate the claim that greater labor market flexibility underlies U.S.-OECD-Europe differences in outcomes; and consider the costs that accompanied American employment expansion. My main claim is that the United States paid for its employment expansion with reduced growth of real wages and productivity rather than with relatively costless flexibility. I find that some aspects of flexibility in relative wage setting helped limit U.S. unemployment while others did not, and argue that the disparate experiences of the United Kingdom and Sweden show that a decentralized labor market is neither necessary nor

sufficient for employment-enhancing wage settlements.

## I. Contrasts in Unemployment/ Employment Experiences

Three fundamental facts underlie the European view of American unemployment: first, the 1980's reversal of the longstanding pattern of higher rates of unemployment in the United States than in OECD-Europe (Figure 1A); second, the growth of employment in the United States, evinced in a rising employment/working age population ratio compared to a declining ratio in OECD-Europe and even more dramatically in employment rates adjusted for the sizeable drop in annual hours per employee in Europe (Figure 1B); and, third, the relatively short duration of unemployment spells in the United States, where incomplete spells have averaged from 12 to 20 weeks compared to several years in many OECD European countries (Figure 1C). While spell lengths differ partly because many U.S. spells end in labor force withdrawal (Kim Clark and Lawrence Summers, 1979) adult male durations are so much longer in Europe than in the United States that this cannot explain the differences (OECD, 1987, table R).

Youth unemployment is also widely judged to be a greater problem in Europe than in the United States, though differences in schooling and student work behavior creates problems in comparisons. In some European countries, such as Italy, Spain, France, and the United Kingdom (but not Germany), the ratio of youth to adult unemployment rates exceeds that for the United States. The duration of unemployment among European youths also tends to be quite long, exceeding durations for the young blacks who bear a disproportionate brunt of U.S. unemploy-

\*Harvard University, Cambridge, MA 02138 and NBER.



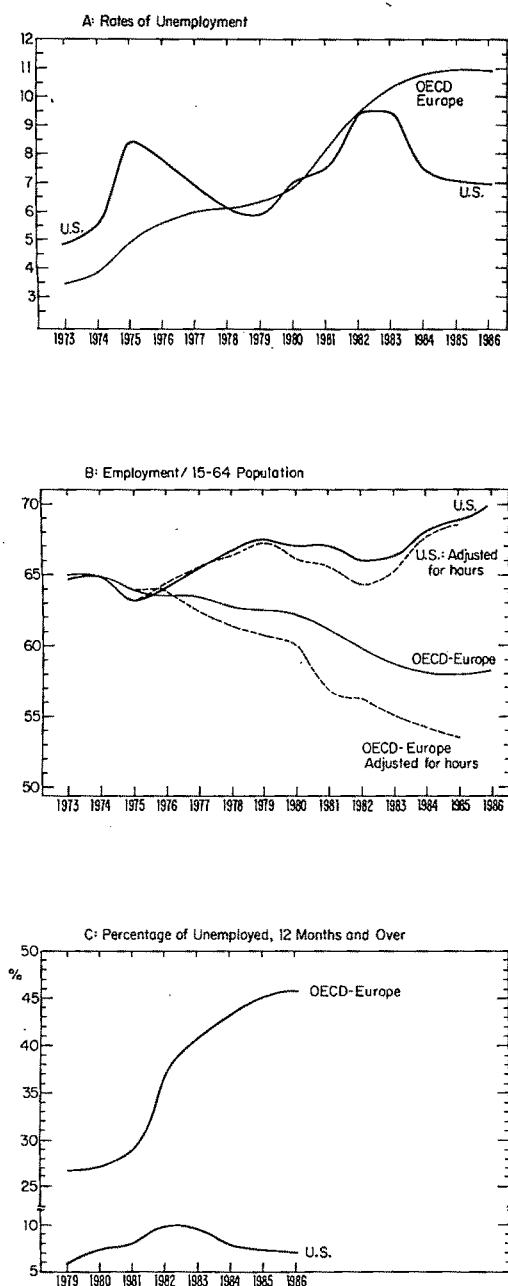


FIGURE 1. U.S.-OECD EUROPEAN EMPLOYMENT AND UNEMPLOYMENT RECORD, 1973-86

Source: OECD (1986a; 1987). OECD-Europe figures in panel B obtained as a weighted average for countries reporting data with 1985 employment used as weights for all years. OECD-Europe figures in panel C obtained as weighted average for all countries reporting data using 1985 unemployment as weights for all years.

ment. And OECD-Europe had nothing like the United States's 1970-80 6 point increase in the employment/population ratio of 16-24-year-olds when the influx of baby boomers into the job market could have created massive youth joblessness.

Less widely recognized, U.S. and European patterns of unemployment also differ along gender lines, with the rate of female unemployment relatively lower in the United States than in Europe (save for the U.K. and Ireland). Because of the differences in youth and female unemployment rates, adult male unemployment rates in the United States are closer to those in Europe than the average rates shown in Figure 1A, offsetting somewhat the presumptively greater cost of unemployment in Europe due to the long durations (OECD, 1987, table 2).

Turning to growth of employment, there is a widespread belief that U.S. growth has been concentrated in low-wage McDonald's-type service jobs. Some hail this as the desirable outcome of flexible wage setting that permits wide variation in pay among industries. Others view it as a sign of American economic decline. In fact, there is nothing special about the growth of service sector employment in the United States. From 1973 to 1984, OECD data (1986a) show that the service share of employment rose by 9 percentage points in OECD-Europe (45 to 54 percent) compared to an increase of 5 points in the United States (63 to 68 percent). Moreover, the shift to services has had only a modest impact on average wages, in part because the service sector includes high-paying professional and business services as well as burger joints. Perhaps most telling, employment and wages have grown more in high-level than low-skill occupations, not what one would expect if the skill structure was deteriorating.

This said, it is the patterns of unemployment and employment shown in Figure 1 that has altered European thinking about the American labor market: "What at the start of the period was being dubbed as a poor labour productivity performance in the United States was being hailed at the end as an impressive job creating performance"

(OECD, 1986b, p. 8). Whereas in the 1950's and 1960's analysts rejected textbook claims that decentralized labor market arrangements work best on the basis of actual outcomes ("you say unfettered labor markets, but our...arbitration tribunals (Australia); bargaining with legal extension (France); shop-floor unionism (U.K.)...produce unemployment far below that in America") in the 1980's the word is flexibility, U.S. style.

So, the question is: to what extent is the U.S. unemployment/employment record the result of the flexibility of decentralized labor markets? The answer turns on the ways in which wage setting and employment determination is in fact more flexible in the United States than in OECD-Europe and on the quantitative contribution of those aspects of flexibility to employment. It requires consideration of aggregate wage change and of wage adjustments and labor mobility along disaggregate industry, occupation, region, etc., lines.

## II. The Contribution of Macro Flexibility

In terms of aggregate wages, recent studies in the Phillips curve tradition suggest that wages in the United States react less to price changes and more to unemployment than wages in many European countries, producing greater "real wage flexibility" (Michael Bruno and Jeffrey Sachs, 1985; David Coe, 1985). As I am uneasy about the robustness of inferences from these time-series regressions, I focus instead on the basic fact that the United States (and Sweden and some other countries) had smaller changes in real wages in the 1970's-early 1980's than most of OECD-Europe and ask: were these modest wage changes (call them the "flexible" response to the post-oil-shock economic world) associated with differences in employment growth across countries? As Table 1 shows, conditional on the growth of real GDP (which increased more rapidly in the United States than in OECD-Europe in total but not in per capita terms), the answer is yes: in each period countries with large increases in real wages (measured by mfg hourly earnings, employee compensation/employees or mfg hourly compensation) had

TABLE 1—REGRESSION COEFFICIENTS  
(AND STANDARD ERRORS) FOR THE  
IMPACT OF REAL WAGES AND OUTPUT ON  
EMPLOYMENT, OECD COUNTRIES, 1960–85

Dependent Variable	Change in	Change in	R <sup>2</sup>
<b>A. Change in ln Employment</b>			
	ln Real Wage	ln GDP	
1960–73	-.57(.11)	.62(.14)	.65
1973–79	-.45(.11)	.71(.17)	.59
1979–84	-.54(.15)	.62(.19)	.56
<b>B. Change in ln Employment</b>			
	ln Real Labor Costs	ln GDP	
1960–73	-.76(.05)	.90(.07)	.94
1973–79	-.62(.10)	.75(.13)	.74
1979–84	-.53(.16)	.88(.22)	.53
<b>C. Comp. Annual Change, Total Mfg Hrs</b>			
	Real Mfg Compensation	Mfg Output	
1960–73	-.53(.08)	.62(.08)	.86
1973–79	-.89(.22)	.36(.22)	.67
1979–85	-.75(.24)	.80(.13)	.81

Source: Panels A and B, 19 OECD countries from London School of Economics Center for Labour Economics-OECD data set. Panel C, 12 Countries (U.S., Canada, Japan, France, Germany, Italy, U.K., Belgium, Denmark, Netherlands, Norway, and Sweden) as given by A. Neef (1986), with wages deflated by GNP deflator, using OECD data.

smaller growth in employment or total hours than countries with small wage increases, with elasticities ranging from  $-0.5$  to  $-0.9$ . As changes in output per worker and wages are highly correlated across countries, moreover, there is a parallel inverse relation between employment and productivity growth, with, for example, wages and productivity growing slowly and employment rapidly in the United States and Sweden and the converse occurring in Belgium, Spain, and in the 1980's the United Kingdom, among others.

If one takes country differences in changes in wages as exogenously determined by labor market institutions (about which I have some doubts) and assumes similar rates of exogenous productivity advance in the United States and OECD-Europe, the estimated wage-employment tradeoff schedule suggests that much of the U.S.-OECD-Europe differences in employment growth was "paid for" by lower real wage growth in the United States. Between 1973 and 1979, for instance, the OECD (1986a) estimates that compound annual rates of GDP growth differed only modestly between the United States and OECD-Europe (2.6 vs. 2.4 percent) while the difference in annual growth in manufactur-

ing wages relative to the GDP deflator was huge (U.S. 1.0 vs. 4.7 percent in OECD-Europe), implying a dominant role for the wage-employment tradeoff in the difference in employment growth. While from 1979 to 1984 differences in real wage growth lessened ( $-0.3$  in U.S. vs.  $0.3$  percent in OECD-Europe) and differences in GDP growth widened ( $2.0$  in U.S. vs.  $1.1$  percent in OECD-Europe) the tradeoff still remains important in the employment story.

Now for problems with this interpretation. First, differences in real wage growth cannot be firmly tied to specific labor market structures. On the one hand, as noted, Sweden and some other countries with quite different labor market institutions than the United States had similar slow real wage growth and sizeable employment expansion, indicating at the minimum that decentralized U.S.-style labor markets were not necessary for real wage moderation (Indeed, the performance of Scandinavia and Austria has fueled claims that corporatist economies perform best in this respect.) On the other hand, the United Kingdom—which has the most decentralized and unregulated labor market in OECD-Europe—had sizeable growth of real wages in the 1980's and experienced the employment consequences thereof. Reinforcing this point, OECD countries with disparate labor market institutions such as Belgium, Australia, and Italy reduced their growth of real wages in the early 1980's, some with noticeable employment consequences, but others with no upswing in employment. Second, workers bargain for money wages while real wages depend on prices as well as wages, raising the possibility that country differences in price setting also contributed to observed differences in real wage patterns across countries. As Robert Solow (1986) has stressed in this context, the tradeoff curve can be interpreted as reflecting the joint determination of wages and employment by exogenous aggregate demand factors, suggesting the need to examine differences in those factors across countries and their relation to the observed changes. Regardless of how one interprets the evidence in Table 1, however, the wage-employment tradeoff represents the key fact

that any explanation of U.S. and OECD-Europe differences must address.

### III. The Contribution of Relative Flexibility

Turning to relative wages and employment (where labor economists feel more comfortable as they can get "in close" to behavior), the evidence suggests that along some dimensions the U.S. labor market has evinced flexibility of the kind likely to be unemployment reducing, while along other dimensions, it has not.

The strongest case for employment-enhancing flexible market responses is found in the changing wage and employment of young workers. Between 1970 and 1983 when baby boomers flooded the U.S. job market, the earnings of the young men fell sharply: real median weekly earnings of workers 16–24 dropped by 25 percent between 1970 and 1985, with the result that the premium of men 25 and older to the 16–24-year-olds jumped from 43 to 90 percent (U.S. Department of Labor). On the demand side, the reduced cost of young workers induced employers to increase the youth share of employment in virtually all industries, from manufacturing to services. In several European countries, by contrast, the relative wages of youths rose or remained steady through the 1970's-early 1980's. Regressions of youth unemployment rates on adult unemployment rates and the ratio of youth to adult pay in a pooled time-series cross section of OECD countries shows that countries where relative pay for youths declined, such as the United States, had less youth unemployment than countries without such responses (David Bloom and myself, 1986; OECD, 1986b). While the drop in youth wages presumably affected overall unemployment more modestly (due to substitution among workers of different ages), it more likely than not dampened total unemployment as well.

Relative wages by region tell a different story. Consider, for example, the summary data on the relation between pay, changes in pay and unemployment across geographic units in countries *X* and *Y* in Table 2: in country *Y*, wages are higher in high unem-

TABLE 2—REGRESSION COEFFICIENTS AND STANDARD ERRORS FOR THE RELATION OF WAGES TO UNEMPLOYMENT, BY AREA

Independent Variables	Dep. Variables:			
	1985-79 Change in ln Wages		1985 Unemployment Rate	
	Country X	Country Y	Country X	Country Y
1985-79 Change in Unemployment Rate	-.92 (.27)	-.43 (.30)		
ln Wage/Earnings, 1985			.03 (.06)	.11 (.02)
Percent employed mfg, 1985	x	x	x	x
Education of workforce, 1985	x	x	x	x
R <sup>2</sup>	.39	.51	.25	.51

Source: My papers (1987, 1988).

Notes: For country X there are 61 areas; the wage change variable is the 1979-85 change in ln weekly wage of male manual workers. For country Y there are 50 areas; the wage change variable is the 1979-85 change in ln hourly mfg earnings.

ployment areas and increases in unemployment only modestly impact wages; in country X, wages are uncorrelated with unemployment at a point in time and declined in areas with relatively rising unemployment—seemingly indicative of a more responsive labor market. Who are the mysterious economies? Country Y is the United States, with states as areas. Country X is the United Kingdom, with counties as areas. While the different pattern of wages might be due to differences in labor market conditions not reflected in unemployment rates, the data seem *prima facie* to reject the notion that geographic wage adjustments are more responsive in unemployment-reducing ways in the United States than in the United Kingdom.

It is not only along geographic dimensions, moreover, that the United States does not seem to have more flexibility in the labor market than other OECD countries. While there are sizeable differences in wage differentials between the United States and some countries (for example, dispersion of industry wages is much smaller in Sweden and Denmark than in the U.S.), analyses of changes in wages by industry in West Germany (Linda Bell, 1986) and the United Kingdom (myself, 1987) show the same factors altering relative wages in those countries with similar magnitudes as in the United

States. In addition, in the 1980's pay differentials by skill and by age changed at least as much in the United Kingdom as in the United States (with no sizeable impact on unemployment).

With respect to mobility, the U.S. labor market evinces enormous short-run changes in employment among establishments, with gross employment flows far exceeding the net flows that determine whether aggregate employment expands or contracts (Jonathan Leonard, 1986). A recent OECD analysis (1987, ch. 4) estimates that the annual rate of "job turnover" (the sum of gross job gains and gross job losses among establishment relative to employment) among Pennsylvania establishments averaged 25.8 percent from 1976 to 1985. If European labor markets were less flexible (say because of hiring and firing laws), one would expect smaller job turnover rates there. But the OECD reports job turnover rates of 23.3 percent in France and 23.5 percent in Sweden. While Germany had a low job turnover rate (16.5 percent), Japan had the lowest (7.7 percent) as well as the lowest unemployment—fair warning to anyone who believes that high mobility is necessary for low unemployment.

#### IV. Costs of Employment Expansion

If the 1970's and 1980's employment growth in the United States resulted from relatively costless flexible labor market adjustments, the European assessment of the experience "as an impressive job creating performance" would be difficult to assail. But the evidence suggests, *au contraire*, that there were substantial costs associated with the U.S. expansion. First, the cross-country analysis of the wage-employment tradeoff suggests that the United States paid for job creation with slow growth in real wages and productivity. The magnitude of the tradeoff was such, moreover, that despite the fact that employment/population rates and annual hours per employee increased in the United States relative to OECD-Europe, *per capita GDP grew at the same 1.3 percent rate*. From this perspective, Americans worked harder for the same gain in living standards

as Europeans. Second, if, as seems reasonable, some persons entered the labor market in response to low earnings of heads of households (for example, married women with children under 1 year of age, whose 1987 participation exceeded 50 percent), their employment reflects a worsening not an improvement in economic well-being. Third, to the extent that the GDP-expansion generated part of employment growth entailed the "double deficits" that turned the nation into the world's greatest debtor, future living standards will be lower, implying an even higher cost to job creation. Finally, even with employment expansion, the U.S. unemployment rate was markedly greater in the 1980's than in the 1970's, which itself exceeded that in the 1960's, while, as noted, unemployment (and wages) were more unequally distributed along some dimensions than in the past.

In sum, the United States paid more for its improved employment and unemployment position relative to OECD-Europe than is recognized by those who peddle flexible decentralized labor markets, U.S. style, as the 1980's Economic Cure-All. There were pluses to the U.S. experience, but there were also costs that make the change in overall economic well-being not so different than in OECD-Europe.

## REFERENCES

- Bell, Linda**, "Wage Rigidity in West Germany: A Comparison with the U.S.," *Federal Reserve Bank of New York Quarterly Review*, Autumn 1986, 11-21.
- Bloom, David and Freeman, Richard**, "The Youth Problem: Age or Generational Crowding," NBR Working Paper No. 1878, 1986.
- Bruno, Michael and Sachs, Jeffrey**, *Economics of Worldwide Stagflation*, Cambridge: Harvard University Press, 1985.
- Clark, Kim and Summers, Lawrence**, "Labor Market Dynamics and Unemployment: A Reconsideration," *Brookings Papers on Economic Activity* 1:1979, 13-60.
- Coe, D. T.**, "Nominal Wages, the NAIRU and Wage Flexibility," *OECD Economic Studies*, Autumn 1985, 87-127.
- Freeman, Richard**, "Are British Wages Unresponsive to Market Forces?," unpublished, London School of Economics, 1987.
- \_\_\_\_\_, "Labor Market Institutions, Constraints, and Performance," *Economic Policy*, forthcoming 1988.
- Leonard, Jonathan**, "On the Size Distribution of Employment and Establishments," NBER Working Paper No. 1951, 1986.
- Neef, A.**, "International Trends in Productivity, Labor Costs in Manufacturing," *Monthly Labor Review*, December 1986, 109, 12-17.
- Solow, Robert**, "Unemployment: Getting the Questions Right," *Economica*, Suppl.: 1986, 53, S23-35.
- Organisation for Economic Cooperation and Development**, (1986a) *Statistiques Retrospectives*, Paris 1986.
- \_\_\_\_\_, (1986b) *Employment Outlook*, Paris, September 1986.
- \_\_\_\_\_, *Employment Outlook*, Paris, September 1987.
- U.S. Department of Labor**, "Median Weekly Earnings of Full-Time Wage and Salary Workers by Selected Characteristics," various editions.

## *ECONOMIC ASPECTS OF PRODUCT LIABILITY*<sup>†</sup>

### Product Liability and Regulation: Establishing the Appropriate Institutional Division of Labor

By W. KIP VISCUSI\*

Society has several institutional mechanisms that promote the control of product health and safety risks and compensation of the income losses that these risks generated. For risks traded in the market, economic forces at work foster each of these objectives. Social insurance programs, such as worker's compensation, promote the compensation objective directly and influence safety incentives through the merit-rating procedure. Two additional institutional mechanisms, which are the focus of this paper, are tort liability and regulation. Each of these institutions has assumed a more active role in the last two decades and has been the focus of considerable academic and policy debate.

What is most noteworthy about these discussions is that both policymakers and economic analysts generally view each institution as the only societal response to the risk. In the field of legal scholarship, this narrow approach has been termed the "tortcentric" perspective by Richard Stewart (1987a, b). Such a piecemeal approach may be necessary in some cases as an analytic convenience, but it neglects potentially important interactions of the two systems. In this paper I explore the nature of the institutional interactions in Section I and examine the ap-

propriate institutional design in Section II. The general conclusion is that risk regulation should play a dominant role in augmenting market incentives for risk reduction and that the scope of product liability remedies should be scaled back to reflect its subsidiary role.

#### I. The Overlap Between Regulation and Product Liability Law

Both risk regulation policies and product liability law have as an objective the control of product safety risks. In the case of risk regulation, neither the general regulatory agencies nor the special mission agencies make any specific allowance for the role played by the tort liability system in promoting safety incentives.

To the extent that product liability lawsuits play a role, it is often the opposite of what is desirable. Prominent lawsuits against products often prompt additional regulation that will bolster the incentives being provided by the courts. In the case of asbestos, for example, the wave of asbestos litigation was followed by tightened OSHA regulation of asbestos, with an average cost per life saved of \$89 million. In addition, EPA has proposed asbestos regulation with a cost of \$104 million per life saved. Rather than substituting for regulation, product liability lawsuits may generate additional regulation.

Compliance with government regulation likewise does not ensure that the product will not be the subject of product liability suits. Regulatory compliance is admissible as a defense, but is not conclusive. For example, the National Traffic and Motor Vehicle Safety Act explicitly states that compliance "does not exempt any person from liability under common law." Regulatory compliance

<sup>†</sup>*Discussants:* Jerome Culp, Duke University; Victor Goldberg, Columbia University and Northwestern University; Robert W. Crandall, The Brookings Institution.

\*Professor of Economics, Department of Economics, Northwestern University, Evanston, IL 60201. This work was undertaken as part of my work in the American Law Institute Project on Compensation and Liability for Product and Process Injuries described more fully in my earlier paper (1987).

is not entirely irrelevant, as companies may introduce evidence of compliance to show that the product has a favorable risk-utility balance and as a consequence should not be considered defective.

Although regulatory compliance at best provides weak support for the product manufacturer's defense, regulatory violations have a much more influential impact in demonstrating manufacturer negligence. Some courts have concluded that such violations constitute evidence of negligence per se. One such instance involved an oral contraceptive manufacturer's failure to include the patient package insert mandated by the FDA. Moreover, it is generally accepted that courts cannot set safety standards lower than those of a legislative body, which all but ensures that product liability and regulatory enforcement sanctions will both be operative for firms that violate regulations. In cases of noncompliance, product liability costs augment the inadequate incentives for compliance created by the regulatory enforcement mechanism. In this class of instances, the institutional mechanisms complement one another.

The extent of the overlap is suggested by data on closed product liability claims presented in Table 1. These breakdowns were generated using Insurance Services Office data on over 10,000 product liability claims closed in 1977, described more fully in my papers (1986, 1988). The first two columns of data are the summary columns pertaining to whether or not the claimant alleged that there were regulatory violations. The final three columns pertain to the type of violations that were alleged: violation of Consumer Product Safety Act (CPSC) standards, violation of Occupational Safety and Health Act (OSHA) standards, or violations of other standards (for example, state, federal, or municipal regulations).

Overall, regulatory violations are cited by claimants in 19 percent of product claims and 28 percent of job-related product liability claims. The expanded scope of government regulations over the past decade no doubt has increased the institutional overlap, which was already substantial a decade ago. Just under half of the violations are for OSHA and CPSC standards, with the mix

TABLE 1—THE EFFECT OF REGULATORY VIOLATIONS ON THE DISPOSITION OF CLAIMS

Variable Category	Violations: Fraction in the Category				
	None	Any	CPSC	OSHA	Other
<b>Product Injuries</b>					
Claims	.81	.19	.06	.02	.11
Successful Claims	.76	.81	.80	.83	.82
Claims Dropped	.20	.13	.13	.13	.13
Settled out of Court	.77	.83	.81	.85	.84
Claimant Wins					
Court Case	.41	.33	.29	.19	.36
<b>On-the-Job Injuries</b>					
Claims	.72	.28	.04	.08	.16
Successful Claims	.60	.72	.66	.71	.74
Claims Dropped	.28	.15	.23	.15	.13
Settled out of Court	.65	.75	.74	.77	.75
Claimant Wins					
Court Case	.25	.40	0	.40	.43

for these two institutions following the expected patterns for job-related and off-the-job injuries.

Regulatory violations enhance the chance of a successful claim, as one might expect given the legal framework that is applicable. For off-the-job injuries, claims with alleged regulatory violations have a 5 percent greater chance of receiving some positive awards and for job-related claims there is a 12 percent differential. This greater effect for job-related claims may be due in part to the greater marginal improvement that is possible for a claims group with a lower rate of success. The success rate for job-related product claims is below that for off-the-job injury since third-party suits are often inappropriate, and are simply used as a means to evade the requirement that workers' compensation be the exclusive remedy against one's employer.

The influence of regulatory violations on the disposition of claims is illustrated by the data in Table 1, as well as by the regression results in Table 2. In each case, the dependent variable was regressed on a constant, the size of the bodily injury loss, and either a dummy variable for any regulatory violation or a series of three dummy variables for whether the violation was for CPSC standards, OSHA standards, or standards of some other governmental body. Since the unit of observation is the individual claim, the dependent variable is a 0-1 dummy vari-

TABLE 2—REGRESSION ESTIMATES OF THE EFFECT OF REGULATORY VIOLATIONS

Dependent Variable	Violations: Coefficients			
	Any	CPSC	OSHA	Other
<b>Product Injuries</b>				
Drop Claim	-0.521 <sup>a</sup> (0.078)	-0.268 <sup>a</sup> (0.102)	-0.248 (0.171)	-0.314 <sup>a</sup> (0.110)
Settle Claim	0.370 <sup>a</sup> (0.070)	0.185 <sup>a</sup> (0.072)	0.135 (0.150)	0.262 <sup>a</sup> (0.099)
Out-of-Court Settlement	0.224 <sup>a</sup> (0.033)	0.304 <sup>a</sup> (0.045)	-0.134 <sup>a</sup> (0.069)	0.070 (0.047)
Claimant Wins Court Case	0.089 (0.227)	0.298 (0.291)	-0.156 (0.490)	-0.060 (0.323)
<b>On-the-Job Injuries</b>				
Drop Claim	-0.803 <sup>a</sup> (0.157)	-0.175 (0.229)	-0.382 (0.214)	-0.646 <sup>a</sup> (0.235)
Settle Claim	0.502 <sup>a</sup> (0.133)	0.248 (0.199)	0.362 <sup>a</sup> (0.186)	0.161 (0.190)
Out-of-Court Settlement	0.219 <sup>a</sup> (0.100)	0.198 (0.145)	0.023 (0.135)	0.131 (0.142)
Claimant Wins Court Case	0.810 <sup>a</sup> (0.345)	-1.516 <sup>a</sup> (0.660)	0.356 (0.463)	1.375 <sup>a</sup> (0.428)

Note: Standard errors are shown in parentheses.

<sup>a</sup>Coefficients that are statistically significant at the 5 percent level, one-tailed test.

able in all but one case, and logit estimation is employed. Standard OLS methods are used for the one continuous variable pertaining to the size of the out-of-court settlement (i.e., the natural logarithm of the bodily injury payment).

The pattern of results in Tables 1 and 2 is quite similar. Claimants will be more reluctant to drop a claim if their probability of success in court is enhanced by a regulatory violation. For product injuries, the drop probability is .07 lower if some regulatory violation has been alleged, and all but the OSHA regulatory violation variable is statistically significant (5 percent level) with the expected sign. For on-the-job injuries, the drop probability difference is .13 when there are regulatory violations, and all but the CPSC regulatory violation variable are negative and statistically significant. The weakness of the OSHA variable for off-the-job injuries and the CPSC variable for job-related injuries is expected given the emphasis of these policies.

The effect of regulatory violations on out-of-court settlements depends on whether it boosts the amounts defendants offer by more than the increase in the claimant's reservation settlement amount. With symmetric payoffs, there will be no effect. Payoff asym-

metry may be introduced if firms will face additional lawsuits involving the product if there is a successful court case against it. Firms will also have relatively higher payoff levels to the extent that claimant risk aversion reduces the certainty equivalent of an expected court award, as in my earlier paper (1988).

The empirical results suggest that regulatory violations are consequential and that they have a relatively greater effect on the willingness of firms to settle such cases than on claimant reservation prices. Out-of-court settlements are 6 percent greater for product claims and 10 percent greater for on-the-job injuries when there are regulatory violations. All of the eight regulatory coefficients have a positive effect on the probability of an out-of-court settlement, with five of them being statistically significant.

Regulatory violations similarly should have a positive effect on the level of out-of-court settlements since settlements should be a weighted average of firms' offer amounts and the reservation settlement level, each of which will be increased by regulatory violations. The aggregative regulatory violation dummy variables perform as expected, but the other more refined variables perform less strongly and in one case, the OSHA variable in the product injury equation, has an unexpected sign.

The final empirical issue—the effect on the probability that a claimant will win a court case—is more difficult to assess since only 4 percent of the claims in the sample reached a court verdict. The only statistically significant effects are the expected positive effects of both the aggregative violation variable and the “other violation” variable and the negative effect of the CPSC variable on claimant success for on-the-job injuries. The unexpected CPSC effect may reflect some omitted aspect of this narrowly defined case group.

Overall, regulatory violations do have a significant effect on the outcome of product liability claims. The direction of the effect follows the pattern one expects for economic variables that enhance the prospects of a claim, with the results following the litigation patterns described in my papers (1986,



1988). Regulatory violations enhance the prospect of a claim's success and appear to affect the firm's expected losses more than the claimant's expected gains. Regulatory violations reduce the probability that a claim will be dropped, increase the likelihood of an out-of-court settlement, increase the size of such settlements, and enhance the claimant's prospects in court actions.

## II. Restructuring the Institutional Interactions

To better promote efficient levels of risk and insurance, I propose the following modification of the tort liability structure. Firms should be exempted from potential liability in court actions if they can demonstrate either compliance with a government regulation that leads to an efficient degree of safety, or the use of a hazard warnings program that leads the market to promote an efficient level of risk. More generally, the risk-utility test applied in product liability tests could be amended to exempt all products for which manufacturers can demonstrate that the risk level is efficient.

Consider first the objective of providing an efficient level of product safety. For products traded in the market, economic forces will be the principal force generating safety incentives for perceived risks. Merit rating for social insurance will also be instrumental for products used in the workplace, as my recent research with Michael Moore (1988) indicates that occupational fatalities would be about 45 percent greater in the absence of workers' compensation.

If these incentives are not adequate, risk-regulation programs that in effect provide a minimum safety constraint are well-suited to the task since these policy mechanisms are targeted explicitly at firms' safety decisions. Most government regulations are designed to promote a level of safety that is more stringent than the economically efficient risk level so that regulatory compliance is often an indication of adequate product safety levels.

Regulatory constraints do not provide any incentives once compliance has been achieved. In terms of institutional overlap, this on-off character of regulatory incentives is one advantage over injury taxes and pollu-

tion tax approaches, since there is no combined effect of regulatory incentives and product liability incentives once compliance at an efficient safety level is reached. For firms out of compliance with the regulation, which is often the case, one can view product liability awards against noncomplying firms as providing an additional compliance incentive. Under the current legal framework, once compliance has been achieved with an adequate standard, firms will face the prospect of additional tort liability. These potential costs will create inefficient incentives for safety, leading firms to produce safety above the level of the regulatory standard.

One cannot rely on tort liability in lieu of regulation since product liability incentives are ill-suited to the task. Not all injured parties file claims, and court awards are far below what is required to promote efficient safety incentives. In the case of fatalities, the courts' valuation of the appropriate compensation for wrongful death is more than an order of magnitude below the value of life that is appropriate from the standpoint of injury prevention. Society should rely on regulation rather than tort liability to address any market failures.

The other policy objective is that of efficient insurance of accident victims. The emergence of the strict liability doctrine was due in large part to a belief by some legal scholars that firms should act as insurers of product losses by incorporating the cost of insurance in the product price and spreading these costs among all consumers.

This approach, while not without superficial appeal, has several shortcomings. First, the rationale was developed before the advent of medicare and medicaid, the increase in workers' compensation benefit levels, and the extensive health and life insurance coverage of the American work force. Since there is generally no offset from product liability awards for social and private insurance coverage, a greater danger than inadequate insurance may be that these awards will lead to overinsurance and an efficiency loss. Second, it is generally inefficient to insure each risk separately on a product-by-product basis. This basic principle of insurance

coverage has been noted since the classic paper by Robert Eisner and Robert Strotz (1961), who observed that consumer's purchase of flight insurance is irrational. Third, the high transactions costs associated with litigation comprise a much greater percentage of compensation than do standard insurance loading costs so that the courts should be viewed as a very inefficient insurer. Fourth, shifting all of the cost of product risks to the manufacturers reduces the consumer's incentive to take care, which may be particularly important when it is property damage rather than one's life that is at risk. Finally, when there are important problems of ascertaining causality, as in the case of toxic hazards, court awards that do not scale the awards based on the product's probabilistic contribution to the adverse outcome will not generate the correct incentives.

Some of these economic issues have been raised with respect to other proposals to deal with the product liability crises. Proposals have been made to cap awards, to abolish strict liability, and to replace the entire tort liability system with an administrative compensation mechanism. My proposal is more limited in that it is only intended to reduce the overlap between regulation and product liability once firms have met an efficient safety standard. The impetus for this proposal is not generated by a desire to reduce the product liability burden but stems from an attempt to establish a coordinated strategy that recognizes the role of the multiple institutions at work. The presence of multiple institutions affecting safety, not just one,

defines the nature of firm's economic environment and should begin to be recognized by economists and legal scholars.

## REFERENCES

- Eisner, Robert and Strotz, Robert, "Flight Insurance and the Theory of Choice," *Journal of Political Economy*, August 1961, 68, 355-69.
- Moore, Michael and Viscusi, W. Kip, "Promoting Safety through Workers' Compensation: The Efficacy and Net Wage Costs of Injury Insurance," Working Paper, Northwestern University, January 1988.
- Stewart, Richard, (1987a) "Compensation and Liability for Product and Process Injuries: Fall 1987 Progress Report," American Law Institute, 1987.
- \_\_\_\_\_, (1987b) "The Roles of Liability and Regulation in Controlling Enterprise Risks," American Law Institute Project on Compensation and Liability for Product and Process Injuries, 1987.
- Viscusi, W. Kip, "The Determinants of the Disposition of Product Liability Claims and Compensation for Bodily Injury," *Journal of Legal Studies*, No. 2, 1986, 15, 321-46.
- \_\_\_\_\_, "Product Liability Litigation with Risk Aversion," *Journal of Legal Studies*, No. 1, 1988.
- \_\_\_\_\_, "Tort Liability and Regulation: The Economic Basis for Assigning Institutional Roles," American Law Institute Project on Compensation and Liability for Product and Process Injuries, 1987.

# The Political Economy of Workers' Compensation: Lessons For Product Liability

By PATRICIA M. DANZON\*

Tort awards for product-related injuries have risen rapidly in recent years. This trend reflects the outcome of court-made decisions, tempered only recently by modest statutory constraints. The workers' compensation (WC) system, under which employers are strictly liable for work-related injuries, is governed entirely by statute at the state level. It provides much lower benefits than does the tort system. There is little presumption that statutory choices for product liability reflect a social optimum, since voters in each state bear a larger share of the costs than the benefits of limiting consumers' rights against product manufacturers, many of whom are located out of state. By contrast, given the standard assumption that the costs and benefits of WC are borne by workers through compensating wage differentials, the WC system provides evidence on collective choices for compensation when costs and benefits of the political choices are internalized within the decision-making jurisdiction. This paper analyzes the political economy of the WC system. The purpose is to investigate whose preferences are reflected in the choice of the WC benefit structure and what lessons can be learned for the optimal design of compensation for product-related injuries and other injuries currently compensated through the tort system.

One caveat is in order. The formal analysis of this paper views WC benefits as designed to provide compensation, thereby ignoring their effect on incentives for care by employers and employees. Ignoring deter-

rence is appropriate only if employee moral hazard is negligible and if compensating wage demands for job risk provide employers with optimal incentives for safety. To the extent deterrence concerns are different in product liability, and if the single tort award must serve the dual function of deterrence and compensation, normative inferences from WC to product liability are tentative.

Previous analyses of WC have tended to conclude that WC benefits provide suboptimal compensation. The policy-oriented literature has long argued that WC benefits are inadequate (for example, *The Report of the National Commission...*, 1972). In a recent study, W. Kip Viscusi and Michael Moore conclude that

the observed rate at which workers are willing to trade off base wage rates for higher levels of compensation greatly exceeds the actuarial rate of trade-off, even taking into account the administrative costs. These results suggest that benefit levels in 1976 were suboptimal, provided that one abstracts from moral hazard considerations.

[1987, p. 260]

Certain features of WC benefits seem inconsistent with basic principles of optimal insurance. Payment is more generous for routine minor injuries than for permanent total disabilities. Some states still limit the duration or the total amount of benefits for permanent disabilities. For temporary and permanent total disabilities, the typical wage replacement rate of two-thirds provides roughly full replacement of after-tax wages (ignoring noncash fringe benefits). But the maximum weekly benefit implies a sharply declining replacement rate at higher wage levels. The mean maximum benefit was .43 of the state average weekly wage in 1965; the mean rose to .81 in 1985, with a range of .36 to 2.32 (see my 1987 paper).

\*Associate Professor, Departments of Health Care Systems and Insurance, Wharton School, University of Pennsylvania, Philadelphia, PA 19104. I am grateful to The Center for Risk and Insurance at the University of Pennsylvania for financial support; to the National Council on Compensation Insurance for data; and to Dong Han Chang for valuable research assistance.

A finding that WC benefit levels are sub-optimal (i.e., below the level that workers would be willing to pay for) would be surprising since it would imply failure to maximize utility of workers and to minimize costs for employers. However the conclusion that WC benefits are suboptimal ignores important differences between WC benefits and the model of the individual demand for compensation that underlies the inference of suboptimality. First, the WC benefit structure, like any social insurance program, is a public good for all individuals covered. With heterogeneity of worker preferences, the common benefit structure is unlikely to be simultaneously optimal for all workers. This raises both the positive question of how the common benefit structure is determined and the normative question of the optimal level of such benefits.

Second, the WC system is only one among several possible sources of insurance for wage loss and medical expense. The choice of state-level public coverages such as WC should be viewed as simultaneously determined with private health and disability insurance, taking as given the structure of federal programs such as SSDI and medicare. The optimal structure of mandated public programs depends on the functioning of markets for private insurance. If private coverages were available at comparable cost to WC, since private coverages can be matched more closely to individual preferences, it would be irrational for a state to incur the deadweight cost of mandating uniform WC benefits. However, if private insurance markets are subject to adverse selection, myopia or free riding, mandatory coverages may be Pareto improving.

### I. Individual Choice of Benefits

Consider first the case where insurance is a pure private good. In each period the worker faces an exogenous probability of injury  $p$ . If no injury occurs he receives a wage  $W$ ; if an injury occurs he receives wage replacement benefits  $K$ . A cost-minimizing employer would select the cash wage  $W$  and benefits  $K$  to maximize the employee's utility, subject to the constraint that the ex-

pected cost of the compensation package is equal to the potential wage with zero benefits  $W^e$ , which is also the value of marginal product under profit maximization. The utility-maximizing level of benefits satisfies

$$(1) \quad U_0 = U_1(1-t)(1+h),$$

where  $t$  is the worker's marginal tax rate;  $h \geq 0$  is the administrative cost per dollar of expected benefits (the load on the employer's insurance); and subscripts 0 and 1 denote the states of injury and no injury, respectively.<sup>1</sup> The individual's preferred replacement rate ( $k = K/W$ ) can be written

$$(2) \quad k^* = k(W^e, t, p, h).$$

Comparative statics analysis indicates that  $k_h^* < 0$ . With state independent utility and decreasing absolute risk aversion,  $k_{W^e}^* < 0$  and  $k_p^* < 0$  if taxes are proportional. With progressive taxes ( $dt/dW^e > 0$ ),  $k_t^*$  would be positive; indeed, if  $(1-t)(1+h) < 1$ , then  $k^*$  would exceed unity. But if there is moral hazard with respect to either the occurrence of injuries or the duration of claims, this would impose the additional constraint  $K \leq W_{(1-t)}$ ; with progressive tax rates this implies  $k_t^* < 0$  and  $k_{W^e}^* < 0$ .<sup>2</sup> However, the very sharply declining replacement rate implied by the maximum benefit is not predicted. These results may not hold if utility is state dependent.

### II. Collective Choice of Benefits

Since the WC benefit structure is a public good within each state, equation (2) cannot be estimated for individuals. Given the small number of states, the assumption of Tiebout sorting of individuals to achieve homogene-

<sup>1</sup>This condition for optimal compensation when wages are taxed is also derived in Viscusi and Moore.

<sup>2</sup>For derivations, see my earlier paper. Although large firms are self-insured or self-rated, perfect experience rating at the firm level may be insufficient to eliminate moral hazard at the level of the individual worker. R. J. Butler and J. D. Worrall (1983) conclude that there is a positive elasticity of claims with respect to benefit levels.

ty of preferences within states is not plausible. Following T. E. Borcherding and R. T. Deacon (1972) and T. Bergstrom and R. Goodman (1973) the choices of benefit levels across states can provide information about individual preferences under certain assumptions, specifically: 1) each voter chooses the  $\bar{k}$  that maximizes his (or her) utility, given his "tax price" ( $s$ ) per unit of  $K$ ; 2) each voter's tax price  $s$  does not vary with the level of  $K$ ; 3) in each state, the quantity supplied is the median quantity demanded, which is the quantity demanded by the individual of median income (i.e., there is no vote trading); and 4) income distributions are proportional, as defined by Bergstrom and Goodman (p. 286). Given these assumptions, each observation is an observation on the demand curve of a consumer with median income given his tax price.

The most difficult variable to measure is the tax price ( $s$ ). For publicly provided services such as education, each voter's tax share is determined by legislation. But, for public goods where the publicness lies in the mandating of a common level of private purchase, each individual's price depends on the prices he faces in private markets. In the short run (with all factors fixed in their current employments), the supply price per unit of  $K$  to the  $i$ th worker in firm  $j$  is simply  $s_{ij} = (1 + h_j)p_j/(1 - p_j)$ , assuming that firms are perfectly experience rated and each worker pays a fully compensating wage offset. Let  $k_{ij}^*$  denote the preferred replacement rate of the  $i$ th worker in firm  $j$ , given  $s_{ij}$ . It is the solution to equation (1) given the short-run supply price.

But, in the long run, the effective supply price of  $K$  to any worker depends on the distribution of preferences of other workers, and on general equilibrium adjustments to the mandated level of benefits. Assume that the state arbitrarily mandates a replacement rate  $\bar{k}$  such that for workers of type  $L$ ,  $\bar{k} > k_l^*$  and for workers of type  $H$ ,  $\bar{k} < k_h^*$ . In the long run, type  $L$  workers who would prefer less than the statutory level of benefits ( $\bar{k} > k_l^*$ ) would not be willing to pay a fully compensating wage offset if they could get  $k_l^*$  in another state or in the uncovered sector of the economy. Similarly, any type  $H$

worker for whom  $\bar{k} < k_h^*$  and who can get  $k_h^*$  elsewhere would require additional wage compensation. Thus mandatory benefits impose a tax on workers for whom  $\bar{k} \neq k^*$ . The tax for the  $i$ th worker is equal to the difference between the cost to the employer and the worker's valuation of benefits:

$$(3) \quad T_{ij} = p_j/(1 - p_j)[(1 + h_h) - U_0/U_1(1 - t_i)](\bar{k} - k_{ij}^*)$$

assuming within-firm homogeneity. The incidence of the tax depends on general equilibrium adjustments in factor and product markets. For any worker, the long-run supply price of  $K$  thus depends his share of the "tax" from mandating nonoptimal benefits for other workers.

For any  $k$ , the effective tax  $T$  is more likely to be positive in small firms, since the load  $h$  is an inverse function of firm size. The magnitude of the tax also depends on the cost of supplementary insurance. Let  $g$  denote the load on private insurance. If  $g < h$  (perfect private supplementation), then  $\bar{k} < k_h^*$  imposes no tax on  $H$ . In practice, the per capita tax from  $\bar{k} \neq k^*$  is likely to be higher for  $\bar{k} > k^*$  than for  $\bar{k} < k^*$ , with some differences by type of benefit. Sick pay and group long-term disability (LTD) insurance are very good substitutes for WC wage replacement for high-wage workers (at least in large firms). Although most lower-wage workers do not have LTD coverage, SSDI provides replacement rates at least equal to the maximum that private insurers would permit. There is no private coverage comparable to the permanent partial wage loss benefits provided by WC. Private group health insurance is a very good substitute for WC medical benefits for disabilities that leave the worker employable. But, if disability leads to loss of employment with access to group benefits, and if the individual does not qualify for medicare, private nongroup health insurance markets provide poor protection against the risk of becoming high risk. Policies are individually underwritten, preexisting conditions are often surcharged or excluded from coverage, and loading charges are typically between .8 and 1 (with

higher loads for policies that guarantee renewability) compared to loads of roughly .25 or less for WC medical benefits. Thus the excess cost of suboptimal WC wage replacement is probably negligible. For long-term medical benefits and permanent partial wage loss, there is less presumption of asymmetry in the per capita excess cost from  $\bar{k} < k^*$  and  $\bar{k} > k^*$ .

If  $g \leq h$  (perfect private supplementation), there would be unanimous choice of  $k = k_1^*$ , unless  $k_1^*$  is influenced by myopia or an intention to free ride. Even if  $g > h$ , type  $H$  workers may nevertheless vote for  $k_1^*$  if  $g$  is still less than their effective supply price of WC benefits, including their share of the tax imposed on type  $L$  workers by mandating  $\bar{k} > k_1^*$ . Thus the effective price to  $H$  of voting for  $\bar{k} > k_1^*$  depends on the magnitude and the incidence of the tax on  $L$ , which depends on elasticities of factor supply, product demand and factor substitution. In general, if type  $L$  workers are mobile and type  $H$  workers are not, type  $H$  workers may bear part of any excess tax on  $L$ .

The incidence of a tax on one type of labor, where the tax rate differs across states can be analyzed using Peter Mieskowski's (1972) general equilibrium analysis of the incidence of the local property tax on reproducible capital.<sup>3</sup> Assume three factors of production: type  $L$  workers for whom  $k_1^* < \bar{k}$ ; type  $H$  workers for whom  $k_h^* > \bar{k}$ ; and capital  $F$  which includes imperfectly mobile factors such as land and small entrepreneurs. All factors are in fixed supply in the aggregate.  $L$  is perfectly mobile among states but  $H$  and  $F$  are imperfectly mobile. If the tax on  $L$  is uniform across states, the full incidence is on  $L$ . But, if the tax rate differs across states,  $L$  in high-tax states will not bear the cost differential in these states since wages of  $L$  ( $W_L$ ) will be equalized in all employments.  $W_L$  falls by the average cost of benefits, including the average tax due to nonoptimal benefits. But the incidence of the

deviations from the mean tax (both positive and negative) is on consumers and other immobile factors. Forward shifting may be possible for nontraded goods such as some retail trade, services, and construction. This is more likely if small firms, that face a relatively high tax rate due to higher costs of providing insurance and safety, do not compete in domestic markets with large firms that face lower loads for insurance and economies of scale in producing safety. There may also be backward shifting to imperfectly mobile factors in high-benefit states and, in particular, to immobile factors in high-cost firms in high-benefit states. Of course, if  $L$  in high-cost firms is imperfectly mobile, then it will bear (part of) the excess tax.

Thus with heterogeneous preferences and a common benefit structure, the standard assumption of an individually actuarially fair compensating wage differential for WC benefits may be incorrect and the choice of WC benefit levels may be affected. If type  $H$  are less mobile than type  $L$  workers, type  $H$  face an increasing marginal cost per unit of  $K$ ,  $h'$ , where  $h'$  is positively related to  $(\bar{k} - k_1^*)$ , to the elasticity of demand for domestically produced goods and to complementarities in production. If  $h' < g$ , there would be unanimous choice of  $k_1^*$  (ignoring myopia and free riding). This choice would be optimal in the sense that it avoids any deadweight loss from imposing a common level of benefits on individuals with heterogeneous preferences. With  $h' > g$  mandatory benefits impose a deadweight loss and there is no presumption that it will be minimized in the aggregate with a median voter model of political choice. However, provided the median voter bears some share of the excess costs imposed on other workers, he would vote for a lower  $\bar{k}$  than if  $\bar{k}$  were a pure private good.

## II. Empirical estimates

Table 1 reports OLS estimates for the log of the maximum weekly benefit ( $MAX$ ) for temporary total and permanent total disability, for approximately 37 states in 1970, 1975, 1980, and 1985.  $MAX$  is a public good for all workers with wages above the threshold

<sup>3</sup>Paul Courant (1977) shows that the Mieskowski model is only approximately correct, but that suffices to establish the point being made here.

TABLE 1—MAXIMUM WEEKLY CASH BENEFIT (LOG)  
(1970, 1975, 1980, 1985)

Variable	Coefficient	t-Statistic
Intercept	4.724	3.52
Wage (LOG)	0.083	0.41
Injury Rate	-0.000	-0.04
POOR <sup>a</sup>	-0.022	-3.15
SMALL <sup>a</sup>	-0.047	-3.38
UNION <sup>a</sup>	0.003	0.60
MANUF <sup>a</sup>	0.004	0.76
AGRIC <sup>a</sup>	0.280	6.38
MINING <sup>a</sup>	0.007	0.36
CONST <sup>a</sup>	0.026	-0.96
SERVICES <sup>a</sup>	0.034	3.56
EDUC > 12	0.007	1.16
D75	0.083	1.30
D80	0.196	2.09
D85	0.153	1.14
R <sup>2</sup>	.739	
n = 146		

<sup>a</sup>Measured as percent.

at which *MAX* is a binding constraint on the replacement rate, but for lower-wage workers it should be irrelevant if compensating wage differentials are individually fair and general equilibrium effects are irrelevant. The significant negative coefficient of the percent of low-income families (*POOR*) is consistent with the hypothesis that general equilibrium effects matter. Benefits are negatively related to the percent of workers in establishments of 20 or fewer employees (*SMALL*) which is consistent with a negative price elasticity of demand. The significant positive coefficients of percent of workers in agriculture (*AGRIC*) and services (*SERVICES*) could reflect the higher cost of private supplementation in these industries, as evidenced by the fact that a disproportionately high percentage of workers in these industries lack private insurance. Dummy variables for 1975, 1980, and 1985 are positive, although not highly significant. This suggests either that the threat of federal intervention following the National Commission had an effect or that WC is subject to some of the same influences that have lead to rising real tort awards and that these influences are not captured by the explanatory variables included here. The income elasticity (*WAGE*) is insignificantly different from zero. Union-

ization and other measures of industrial mix are also insignificant.

#### IV. Conclusions

This analysis has several implications for interpreting the choice of WC benefits and drawing inferences for product liability. First, no worker votes for less than the benefits he (or she) is willing to pay for, given the effective supply price,  $h'$ . But  $h'$  depends not only on the load on his own employer's insurance but also of the difference between his preferences and those of other workers. General equilibrium adjustments in labor and product markets internalize to some extent to each worker the excess costs that his choices impose on other workers. Second, willingness to pay for state-level mandatory benefits also depends on the availability of private supplementary benefits and federally financed public programs. Thus for some workers WC benefits may appear to be sub-optimal; but this is true only ignoring supplementation and ignoring the deadweight costs imposed on other workers from mandating higher benefits.

Both factors—supplementation and deadweight costs from imposing common benefits on heterogeneous individuals—apply equally in the case of insurance for product-related injuries. WC benefits therefore provide a reasonable guide for optimal compensation through the tort system, ignoring deterrence.

#### REFERENCES

- Bergstrom, T. and Goodman, R., "Private Demands for Public Goods," *American Economic Review*, June 1973, 63, 286-96.
- Borcherding, T. E. and Deacon, R. T., "The Demand for the Services of Nonfederal Governments," *American Economic Review*, December 1972, 62, 891-901.
- Butler, R. J. and Worrall, J. D., "Workers' Compensation: Benefit and Injury Claim Rates in the Seventies," *Review of Economics and Statistics*, November 1983, 4, 580-99.
- Courant, P. N., "A General Equilibrium Model of Heterogeneous Local Property Taxes,"

- Journal of Public Economics*, December 1977, 8, 313-27.
- Danzon, Patricia M., "Determinants of Workers' Compensation Benefit Levels," paper presented at the Seventh Annual Seminar on Economic Issues in Workers' Compensation, November 1987.
- Mieskowski, P., "The Property Tax: An Excise Tax or a Profits' Tax?," *Journal of Public Economics*, April 1972, 1, 73-96.
- Viscusi, W. K. and Moore, M. J., "Workers' Compensation: Wage Effects, Benefit Inadequacies, and the Value of Health Losses," *Review of Economics and Statistics*, May 1987, 69, 249-261.
- The Report of the National Commission on State Workmen's Compensation Laws*, Washington, 1972.



# The Political Economy of Product Liability Reform

By RICHARD A. EPSTEIN\*

The object of this paper is to offer some partial explanation of the failure to obtain legislative reform of the product liability system, both at the state and the federal level. In undertaking this inquiry, I believe that some legislative reform is needed, even if much existing legislative reform is misguided.

## I. The Gains from Legislative Reform

The present body of product liability rules has been fashioned by common law judges, with an occasional assist from legislatures. The key premise of the present system is that freedom of contract, even subject to the usual caveats of force and misrepresentation, has no place to play within the product liability system. That assumption makes good sense in the few cases in which defective products are responsible for injuries to bystanders. But it is far more dubious when suits are brought by injured product purchasers or users. To be sure, there is typically no direct contract link between the manufacturer and ultimate purchaser. The absence of explicit contracts, however, is itself a reflection of the present body of legal rules that regulate product use. These rules refuse to enforce a limitation on liability or damages created by contract, and occasionally impose punitive damages on firms bold enough to seek contractual protection. If the law were otherwise, then a manufacturer could—and would—designate a retailer as his agent so as to secure a direct contract with a purchaser and, where possible, a product user. Contractual quiescence today is not a sign of satisfaction with the legal rules as ideal default provisions. Nor is it evidence that transactions costs are so high that voluntary agreements cannot be formed. It is only proof that any effort to contract out of the present

tort system has been effectively blocked by the judges and legislatures who have created the modern law.

The present regime of legal rules has powerful social consequences. The conventional economic assumption is that rational, self-interested parties only enter into contracts that *ex ante* assure them of some joint gain. By that standard, the present system of public regulation of contract terms, whether by common law decisions or statute, imposes significant losses on both sides. Legislation to remove these barriers to private contract could generate some substantial overall gains. Whenever legislation could generate some allocative improvement, it should be possible to divide the gains thereby generated to leave all interested parties better off than before. If the costs of passing legislation and dividing the gains were zero, then inefficient substantive rules would never remain on the public books. Today's product liability rules do not fade, but grow stronger with each passing year. The remainder of this paper helps to explain why this trend is likely to continue. The barriers to needed legislative reforms are simply too great.

## II. The Obstacles to Legislative Reform

In isolating the barriers to legislation, three points are of some importance. First, there is no ideal forum state or federal for product liability legislation. Second, lawyers from *both* the plaintiff and the defendant bar are ideally positioned to block such legislation, and have been able to do so. Third, serious conflict of interests within the class of manufacturers who might benefit from product liability reform legislation hampers their unified efforts.

### A. Where Legislate Reform?

One problem of no little importance is that products are typically sold in national (or international) markets, but are regulated

\*James Parker Hall Professor of Law, University of Chicago, Chicago, IL 60637.

by tort law generated at the state level. About ten years ago I worked on a proposal for legislative reform prepared by the American Insurance Association. When I testified on behalf of that package at hearings before the California State legislature, the first question I was asked had little to do with its merits: "Why should this state pass laws which benefit out of state manufacturers at the expense of in state consumers?" The point of the question was clear, even if the interest-group analysis that lies behind it may be quite complex. Some local interests, such as distributors of out-of-state products, lose from bad product liability laws, but these parties are apt to be outnumbered by the local interests (including local lawyers) who benefit from the present legal regime. So long as legislators can manage their sums, product liability laws, bad from the point of view of the nation at large, may be good from the point of view of a winning coalition of citizens in any particular state. The losses that product liability law imposes upon firms outside the state will not be registered, or at least registered with equal intensity, with the gains to in-state persons. The mismatch between economic interest and voting power leads to an externalization of loss through the political process. The state law product liability rules may reduce overall national wealth, but increase wealth within the state. The political equilibrium seems stable, and suboptimal.

Product manufacturers often have tried to escape this vulnerable situation by striking deals with local groups that also have been heavily hurt by the changes in liability rules. Local governments, and physicians and hospitals are the key players here. The underlying sources of expanded liability are the same, with the decline of customary standards of negligence and the rise of *ad hoc*, open-ended cost-benefit liability formulas. But the coalition of defendants has rarely held. Local governments and local physicians do not have to face the question, "why benefit out-of-state providers at the expense of in-state consumers?" Everyone is in-state. What has happened, therefore, has been a spate of reform on issues, leaving out the product manufacturers and sellers.

Frustrated by reform at the state level, product manufacturers have sought to press for reform at a level where both prospective manufacturers and consumers have equal voice. In practice, there are two possible ways to proceed, and both have been tried without success. The first is to secure some form of interstate compact by which all states agree to adopt some uniform rules that register the preferences of all persons within the country. On its face, the political obstacles to such a master agreement seem formidable, but they have been overcome. The Uniform Commercial Code, which regulates sales, negotiable instruments, and security transactions, has been adopted, virtually in toto, by statute, mainly in the 1950's and 1960's, in all 50 states. Twenty-five years later, the U.C.C. and its body of case law remain remarkably cohesive across state lines, and freedom of contract principles are still dominant on most critical areas. (Consumer transactions, largely outside the code, are another story.) There is simply no powerful interest group that gains from clumsy rules of sales, checks, and security. The insatiable demand for special interest legislation manifests itself in other arenas, such as the battles over interstate branch banking.

During the late 1970's, efforts were made to generate a Uniform Product Liability Law. The Interagency Task Force (from the Departments of Commerce, Labor, and Treasury) put out massive studies of product liability reform, and presented for public discussion many different versions of model reform statutes. But it all quickly went nowhere. While it is easy to see how large banks and other commercial players could benefit from a nationwide free market, it is much harder to make the same calculation with respect to product liability rules. The political forces that were strong enough to stop or dilute isolated initiatives at the state level were strong enough to defeat the uniform proposals, even though they had some modest federal backing.

With the failure of the uniform state approach, matters then moved into the federal arena, where manufacturers were roughly at equal strength with consumer interests. Nonetheless, this strategy has also failed,

and for compelling reasons. The question of tort law in general has long been thought to be a state issue. Even though the Congress today has the undoubted constitutional power to impose a uniform product liability law, its respect for state independence has led it to refuse to act on product liability reform, just as it refused to intervene on automobile no-fault and workers' compensation laws. The shift to the federal forum had the consequence of bringing in new players to the debate, those who cared about federalism even if they did not care about product liability reform. The advocates of comprehensive reform have thus far been beaten back on a collateral set of issues. The federal courts, for example, had little desire to cope with enlarged dockets of product liability cases that a federal product liability law might generate.

### B. *The Lawyers*

Thus far I have written as though the question of product liability reform pits manufacturer against consumer groups. That observation ignores the powerful role of lawyers in reform efforts at both the state and the federal level. Obviously, the plaintiff's bar has a vital interest in preserving that system of laws which maximizes its own welfare. Less obviously, perhaps, the defendant's bar has closely parallel interests. No defendant lawyer has ever made substantial sums of money by being able to win a summary judgment (i.e., judgment without the need for trial) for its clients. The plaintiff's lawyer will know of the probable fate of the suit, which will therefore never be brought because it has no chance of success. Clear, decisive rules that demarcate some zone of conduct in which the manufacturer or seller is *not* liable generates no income to any member on either side of the bar. (The defendant's antitrust bar has also suffered greatly from the substantive revolution in antitrust.) It is therefore in the interest of defendant firms to have a pro-plaintiff set of rules, which makes their own defensive efforts worthwhile for the manufacturers that hire them. Even though inside corporate counsel will disagree, defendant law firms have little

desire to see clear principles of law emerge, for it is so much more remunerative to win after an exhausting slugfest on the facts.

The question is, what is the optimal level of complexity of the rules for the lawyers. They need some interior solution. Neither plaintiff nor defendant lawyers want a set of rules so complex that no lawsuit will be brought at all, just as they do not want a set of rules so simple that their services could be dispensed with in settling cases—that was of course one of the original, if unattained, reform motives behind the workers' compensation statutes. The predicted outcome is a set of rules of genuine complexity that allows both sides of the bar to maximize their expected income, measured as the product of the frequency and expense of lawsuits. That is just what the state common law judges have provided us today. The dominant rules of liability work on the so-called risk-utility test, that asks the jury to decide whether the overall social benefits of marketing or using a given product outweigh their overall costs. There is a large collection of relevant factors—the cost of using alternative designs, the plaintiff's knowledge of the risk, the availability of insurance—that are relevant to the overall outcome, but no matter how combined, these factors yield no determinate rule for any individual case. Discretion is king, and the services of expert lawyers on both sides are indispensable for any party—plaintiff or defendant—to navigate the legal waters.

That dominance carries over from the court room to the legislature, where lawyers are well-positioned to block any legislation. Tort reform usually must begin in the judiciary committees, where lawyers dominate the membership. In addition, the plaintiffs' bar is well organized at the national level and in every state in the union. It has little difficulty in maintaining a united front among its members because it adopts the strong uniform line that all matters of tort law should be decided by the common law courts, which have decided these questions since the time of William the Conqueror.

Given its economic interests, the defendants' bar attacks some of the excesses of the system, but acknowledges its basic

soundness. Its posture, therefore, is to note that no manufacturer need fear liability if well represented at trial, and then to plead faintly on behalf some modest reform. But it will not support any movement toward contract, or any other important structural reform, such as treating compliance with comprehensive federal standards (especially for drugs and automobiles) as dispositive on questions of liability. Lawyers as a group are champions of the status quo.

### *C. The Manufacturers Coalition*

There is a third problem which further hampers the general movement toward product liability reform. There is no cohesion among the class of probable beneficiaries of that reform. As noted above, the plaintiffs' bar has an easy strategy to keep itself together: oppose all reforms. In general, it is hard to drive a wedge in that alliance. Initially, it is easy for the lawyers to observe whether their lobbyists are complying with the general charge. It is also in the interest of most lawyers within the group to want the total prohibition on reform. Individual lawyers tend to specialize, perhaps in drug or machine tool cases. But plaintiffs' lawyers tend also to work in firms, usually small firms, that have experts in different kinds of lawsuits. These firms offer forms of insurance for their members, so they will tend to work to see that all parts of the business are maintained. They have a high level of cohesion.

There is no parallel identity of interest on the manufacturer's side of the reform question. Products liability law is an extremely complex and diffuse body of law. An apricot with a pit inside is a product, but so too is a complex nuclear reactor. The type of reforms that will benefit some product manufacturers and sellers are often of no concern, or less concern, for others. A couple of illustrations shows how far the interests on this side of the struggle diverge.

One key question in product liability law concerns the coordination of workers compensation benefits with the tort actions brought against persons other than the employer, here the manufacturer. The precise

method of coordination is of great concern for machine tool manufacturers, whose products often result in workplace accidents that could well require both employer and manufacturer to chip in to the plaintiff's recovery. The reforms proposed usually try to reduce the manufacturer's contribution. This entire question, however, is of no concern to automobile and drug manufacturers, who are not involved in workplace accidents. When this issue comes on the table, the full force of the plaintiff's bar is against legislative change, but a large fraction of the manufacturers rightly have no interest in the issue at all.

Turnabout is fair play. Drug manufacturers are generally concerned with the adequacy of warning questions and the role of the FDA in setting standards. These are questions for which machine tool manufacturers have no concern at all. Automobile manufacturers are generally preoccupied with the problems of design defect, and have far less worries with the warning questions that are the bane of the drug companies. And so it goes.

Faced with these differences in demands, the proponents of reform face a genuine quandary on how to proceed. The members of the separate industries could get together in the effort to pursue an omnibus package with something for everyone. But it is often beyond their ability to get it all. Legislation is often a matter of horse trades. So when someone proposes a change in the warning provisions, the drug companies will be vitally affected but the other members of their coalition will not. It is hard, therefore, to keep the united front together because bills are constantly revised while winding their way through the legislative process. With each new wrinkle, the coalition has to decide whether to accept, reject, or make a counterproposal. Each trade association (indeed, each individual manufacturer) continually must reassess whether it wants to stay behind the collective reform project as its hue and coloration change. There are no long-term binding contracts. Usually the alliance cannot survive the pounding.

The alternative is for specific-interest groups to seek legislation that is tailored to

their specific industries. They still face the united wrath of the plaintiffs' bar, but they can no longer count on the support of other manufacturers. Unless therefore some strong public outcry is raised, the reform effort will usually fail because it cannot find the winning coalition at any level. The dilemma is complete. Coordinated action across industry groups is not feasible. But individual groups acting alone lack sufficient clout.

### III. Conclusion

The difficulties associated with product liability reform are not likely to vanish with

the next election or change in political fortune, because the structural obstacles to reform seem a constant in our political process. Accordingly, there is a tendency for judge-made common law rules to be the last as well as the first word on many critical issues of tort liability. It is, of course, proper in a jurisprudential sense to say that common law must yield to legislation that falls within proper constitutional limits. But, in practice, the default rules of the common law courts are apt to be a permanent part of the legal landscape. The benefit of good judicial decisions is thereby increased. But so too the cost of bad ones.

## SURPRISES FROM DEREGULATION<sup>†</sup>

### Surprises of Airline Deregulation

By ALFRED E. KAHN\*

Surprises are a product of mistaken expectations and unforeseen outcomes. As for the former, I have no taste for the task of putting together a fair composite depiction of the expectations of the airline deregulation advocates; the fact that they ranged from Ralph Nader to the National Association of Manufacturers suggests how difficult that would be. I will therefore confine this account to my own expectations—and inevitably succumb to the temptation to deploy the evidence selectively, so as to demonstrate my prescience about both the good results and the unpleasant ones.

The subject is irresistible, however, partly because the aboriginal opponents of deregulation have been assembling collages of predictions by the proponents and depictions of the results that, even if authentic in their several parts, turn out to be caricatures in their composite.

The main more or less unpleasant surprises—be assured I will conclude with a brief but heartfelt summary of the pleasant ones—fall under four headings: 1) the turbulence and painfulness of the process; 2) the reconcentration of the industry; 3) the intensification of price discrimination and monopolistic exploitation; and 4) the deterioration in quality of airline service.

#### I. Turmoil

While the advocates of deregulation recognized that competitive markets are inherent-

ly more messy and unstable than tightly regulated ones (see my 1971 study, pp. 12–13, 325–26), and recognized also that radical changes were likely to follow removal of the pervasive restrictions that had been imposed on the industry over the preceding forty years, I doubt that most of us were fully prepared for the explosion of entry, massive restructurings of routes, price wars, labor-management conflict, bankruptcies and consolidations and the generally dismal profit record of the last ten years.<sup>1</sup>

During the period of rapid deregulation, I scoffed at what

seemed to be a general belief among defenders of the present regulatory regime that there is something about airlines that drives businessmen crazy—that once the CAB removes its body from the threshold, they will rush into markets pell-mell, en masse, without regard to the size of each, how many sellers it can sustain, and how many others may be entering at the same time.<sup>2</sup>

I was wrong—at least temporarily—but almost certainly will prove decreasingly so as time goes on.

What inferences are we to draw, however, from these particular surprises?

The turbulent entry of new, much lower-cost carriers, and their ability to quote

<sup>†</sup>*Discussants:* George Kaufman, Loyola University of Chicago; Marvin Koters, American Enterprise Institute; Paul W. MacAvoy, University of Rochester.

\*Robert Julius Thorne Professor of Political Economy, Cornell University, Ithaca, NY 14853 and Special Consultant to National Economic Research Associates, Inc.

<sup>1</sup>The industry's profit margin averaged only 1.30 percent in 1970–77, which was bad enough compared with industry generally, but fell to a puny 0.10 in the 1979–86 period. (Calculations from the Air Transport Association, 1975–87.)

<sup>2</sup>"I cannot believe, in any event, that it requires governmentally-imposed cartelization to make this or any other industry creditworthy" (myself, 1978a, pp. 15–16, 28.)

much lower fares than the incumbents—typically across-the-board—were a clear reflection of the extent to which the latter's costs had become inflated behind the protective wall of regulation, and an illustration of competition doing exactly what we hoped and expected it to do.

Considering the maniacally detailed restrictions on the operating authorities of airline companies under regulation, it would have been shocking if their removal had *not* resulted in a massive reordering of routes: what better proof could there be of the gross inefficiencies engendered by regulation?<sup>3</sup>

Of much greater significance than the changes in the operations of individual companies has been the continuity and expansion of service in the aggregate. Thanks partly to the Essential Air Services Program incorporated in the 1978 Act, not a single community that enjoyed a minimum level of certificated service at the time of deregulation has lost it. Many communities have lost uncertificated service since that date, just as many had under regulation, but that had little or nothing to do with regulation or deregulation (U.S. GAO, 1985, p. 29). The smallest towns, the so-called nonhubs, have as a group experienced practically no change in their average weekly departures between 1978 and 1987, while the small hubs have enjoyed a 42 percent increase (Melvin Brenner, 1988, Figure 15; also myself, 1988b).

The industry's severe financial losses in 1981–83 were the result primarily—perhaps entirely<sup>4</sup>—of the severe recession, the fuel price explosion of 1979–81, and the air traffic controllers strike. (On the other hand, the

very poor financial showing in 1986, a year of general economic prosperity, must be attributed preponderantly to the intense price competition that deregulation unleashed.)

While the industry's return on equity has plummeted almost to zero, its average returns on total invested capital have been no lower since 1978 than before (means shown with standard deviations following in parentheses): 1965–77, 6.3(3.5); 1970–77, 5.3(3.1); 1978–86, 7.2(3.1); 1979–86, 6.4(2.4). (Calculated from Air Transport Association; see also myself, 1988b.) Perhaps equally striking, the volatility of these returns has not increased.

The opponents of deregulation claim that what both they and investment analysts generally see as the perverse tendency of the industry to continue to add to capacity in the face of these poor financial results proves they were right in predicting that unregulated competition would tend chronically to be destructive. The ultimate public concern about the possibility of destructive competition, however, is that it may result in an impairment in the ability of an industry to finance needed expansions of capacity, and a consequent deterioration in service (myself, 1971, pp. 175–76). The triumphant assertions of the critics, therefore, are in effect a concession that this particular threat to the consumer has not in fact materialized—partly, no doubt, because several of the airline companies have been doing very well indeed.

Labor unrest and the insecurity and downward pressure on the wages of the pre-existing labor force have been an undeniable cost of deregulation. From the standpoint of the public, however, grossly monopolistic wage levels are no more acceptable than monopoly profits. The fact that these costs have been unusually severe may just as logically be blamed on the regulation that created vested interests in its perpetuation as on deregulation.

Total employment in the industry actually increased 39 percent between 1976 and 1986. The increase in revenue passenger enplanements by 87 percent during the same decade, and the increase in productivity reflected in these comparative changes are

<sup>3</sup>It was precisely in recognition of the size of the resulting distortions and the unfitness for competitive survival of companies that had been nurtured in a regulatory hothouse for the preceding forty years that I attempted—unsuccessfully—to give the industry time to adjust, by deregulating only gradually. See my 1978b statement, pp. 5–13.

<sup>4</sup>Steven Morrison and Clifford Winston (1986, p. 40) and John Meyer, Clinton Oster, and John Strong (1987, pp. 21–32) both conclude that during the 1980–82 period the financial showing of the industry might have been even worse had it not been deregulated.

among the most important benefits of deregulation.

## II. Reconcentration and the Attenuation of Competition

Just as one of the most pleasant surprises of the early deregulation experience was the large-scale entry of new, highly competitive carriers, so probably the most unpleasant one has been the reversal of that trend—the departures of almost all of them, the reconcentration of the industry both nationally (Brenner, Figure 3) and at the major hubs (Julius Maldutis, 1987, pp. 6–9), the diminishing disciplinary effectiveness of potential entry by totally new firms, and the increased likelihood, in consequence, of monopolistic exploitation. The reasons for these developments are generally familiar and in any event have been thoroughly expounded by Michael Levine (1987)—the advantages of controlled traffic feed, particularly by developing and dominating hubs; the difficulty of rivals mounting an effective challenge at those hubs; the advantages conveyed by ownership of computerized reservations systems (CRS) and frequent flyer programs; the discovery by the incumbents of the superior competitive attractiveness of deeply discounted fares—far lower than their smaller, lower-cost competitors were able to match on an across-the-board basis—targeted (with the help of increasingly sophisticated computerized scheduling) for seats that would otherwise be likely to go out empty; and the flood of mergers and operating agreements between competitors and potential competitors.

Were these developments surprises? Yes, to a large extent. We advocates of deregulation were misled by the apparent lack of evidence of economies of scale—the principal explanation of the differences in cost among the carriers appeared to be differences in their route structures, which we hoped to eliminate by permitting totally free entry and exit—and by the physical mobility of aircraft, which caused us to underestimate the other obstacles to entry. While recognizing the competitive advantages of controlled

traffic feed, we were, as it turned out, overly impressed by the apparently equally great competitive opportunities for specialized turnaround service, and therefore did not anticipate the thoroughgoing movement to hub-and-spoke operations and the dominant role it would play in determining the balance of competitive advantage and disadvantage.

At the same time:

As I specifically observed (1978a, pp. 18–22, and 1979, pp. 5–6), if it was impossible for government officials to predict what kind of route structures would prove ultimately to be the most effective, that was an argument not for perpetuating ignorant regulation but for leaving the decision to the competitive market.

Whatever misgivings one may have about this kind of competition-by-preemption of traffic (and I have more than most economists) one must recognize that the critical advantages of hub-and-spoke operations reflect genuine efficiencies: the superior quality of on-line service (in which passengers change planes from one flight to another of the same carrier) over interline, fuller utilization of larger planes and the possibility of offering a wider range of destinations from all originating points—the principal source, according to Morrison and Winston (pp. 31–33), of the multibillion dollar annual benefit to the flying public attributable to deregulation.

The radical transformation of the operations of the incumbent carriers that enabled them so quickly to overcome the competitive threat of the new entrants was, in very large measure, the beneficent consequence of competition: the successful ones cut their costs, rationalized their route structures, developed extraordinarily efficient CRSs and learned to offer deep discounts to fill their planes.

The concentration process reflected also what many of the advocates of deregulation would characterize as a lamentable failure of the administration to enforce the policies of the antitrust laws—to disallow a single merger or to press for divestiture of the computerized reservation systems or attack a single case of predation. None of these cases would have been easy. All of the mergers, it



could be argued, gave birth to more effective competitors; the harmful effects on competition of major carriers owning CRSs, on the one hand, and the feasibility and desirability of their divestiture, on the other, remain intensely contested; and the feasibility of identifying and moving against instances of predation are extremely uncertain. At the same time, I take perverse satisfaction in having predicted the demise of price-cutting competitors like World and Capitol Airways if we did nothing to limit the predictable geographically discriminatory response of the incumbent carriers to their entry, and in having rejected the conventional wisdom that predation would not pay because any attempt to raise fares after the departure of the price-cutting newcomers would elicit instantaneous competitive reentry.<sup>5</sup>

Despite the now markedly higher concentration of the industry at the national level, it is not at all clear that concentration has gone up in the economically pertinent markets—namely, individual routes. On the contrary, it *appears* that the average number of carriers per route is still higher today than it was under regulation (for a survey of the incomplete evidence, see my 1988b paper).

The relatively small number of airlines were under regulation prevented for the most part from competing with one another; since deregulation they have been free to invade one another's markets, offering whatever combinations of price and service they choose, and they have done so, vigorously.

While, therefore, travelers on flights originating and terminating at the concentrated hubs probably face fewer alternatives now than before deregulation, competition on longer, connecting flights over various hubs has clearly intensified: a Boston/Phoenix passenger, for example, has the choice of nine hubs at which to make connections (Maldutis, p. 9).

The industry remains to this very day far more intensely competitive than it was

before 1978. The opponents of deregulation cannot have it both ways—asserting on the one hand that competition has proved to be a lost cause and, on the other, that it has been and remains catastrophically destructive. They will undoubtedly retort that the process of competition killing itself off is still incomplete. The response—now, as ten years ago—is that the possibility, which no one can deny with total certainty, that competition *may* one day prove not to be viable is hardly a reason to have suppressed it thoroughly in the first place.

### III. Price Discrimination and Monopoly Exploitation

The benefits of price competition under deregulation have been very widespread. Between 1976 (the last year before the CAB began to permit widespread discounting) and 1986, average yields per mile dropped 28.5 percent in real terms. According to the Air Transport Association (1987, p. 5), 90 percent of all passengers in 1986 traveled on discount tickets, at an average 61 percent below coach fare. And while this means that the coach fares themselves have become increasingly fictional, the studies by the Meyer-Oster group show that they have not risen egregiously compared with the levels at which they would have been set under regulation (Meyer et al., pp. 112–13 and 121–22).

The very low fare levels of 1986 and early 1987, reflecting severe price wars, were not sustainable—the industry as a whole lost money—and yields have in the last months of 1987 almost regained 1985 levels.<sup>6</sup> But the decline from 1976 to those 1985 levels would still have represented savings of \$11 billion to airline passengers in 1986 alone.

At the same time, the pressures and benefits of price competition have been unevenly distributed geographically. The troublesome disparities that have emerged are not, however, wholly discriminatory: it costs more to provide service on small airplanes, on thin routes, with the frequency required to meet

<sup>5</sup>Large portions of the memorandum to my fellow CAB members in which I expressed these opinions are reproduced in my 1988a paper.

<sup>6</sup>Information from the Air Transport Association.

the needs of business travelers, than it costs on the dense routes and to serve vacationers.

It is by no means obvious to what extent travelers in the less competitive markets have actually been exploited. What is extremely dubious is that, as is widely assumed, their fares have gone up *because* fares have declined, dramatically, in the more competitive markets—that is to say, that passengers in the thin markets are “subsidizing” the bargains in the dense ones. Such contentions assume that businesses would, irrationally, sell some services for substantial periods of time at prices below incremental costs and others at prices below profit-maximizing levels, raising the latter only after and because competition had forced them to reduce the former. On the contrary, if the introduction of intense price competition on the dense routes has had any effect on prices on the thin ones, it is more likely to have been to reduce than to increase them, because of the ability of many travelers to rearrange their routing to take advantage of the discounts.

The persistence—indeed, intensification—of price discrimination has been a surprise. While I pointed out (1978a, pp. 24–27; 1978c, pp. 39–40, 50–57) that the structure of airline costs—the inevitability and desirability (on quality grounds) of average load factors far below the 100 percent level, with the consequent availability of zero marginal cost seats—clearly suggested that widespread price discrimination would continue under competition, I was at other times so carried away by witnessing the introduction of across-the-board, nondiscriminatory low fares as to predict that, with competition increasingly pervasive, “much of the price discrimination will tend to disappear” (1979, pp. 11–12). I should have recognized that the naturally monopolistic or oligopolistic character of most airline markets (which I had myself observed, 1978a, p. 24) and the inevitable continuation of short-run marginal costs approximating zero promised that these discriminations would continue—indeed, expand—under deregulation.<sup>7</sup>

<sup>7</sup>My colleague Robert Frank recalled this elementary principle to me and its pertinence in the airline context.

Competition in the real world is, inevitably, imperfect. The question before us in 1977–78 was whether the imperfections would be so severe as to justify continuation of the kind of regulation we had practiced in the airline industry in the preceding forty years, at costs to the economy of billions of dollars a year. At the worst, we might now decide that competition is so insufficiently protective of consumer interests on particular routes as to require us to reimpose price *ceilings* in those instances, although the practical difficulties would be enormous. In view of the CAB’s advocacy of a continuation of such ceilings in markets dominated by a single carrier (myself, 1978b; 1978c, p. 46), I hope I do not shock anybody by observing that I probably would have been very reluctant to abandon price ceilings entirely had I had the choice. All the studies of airline pricing since deregulation confirm that reluctance: market concentration does matter; and their general trend over time has been toward the conclusion that it matters a great deal (compare Elizabeth Bailey et al., 1985, p. 199, with Gloria Hurdle et al., 1987, p. 16).

#### IV. Congestion and Delay

Most of us probably did not foresee the deterioration in the average quality of the flying experience, and in particular the congestion and delays that have plagued air travelers in recent years. Fortunately, an audience of economists will readily understand how little this failure constitutes a legitimate criticism of deregulation:

To some considerable extent, these discomforts are a sign of the success of deregulation, not its failure, resulting as they have from the enormous response of travelers to the offer of very low fares for necessarily correspondingly lower-quality service—narrower seating, longer lines, fewer amenities.

The consequent similar deterioration in the quality of service enjoyed by full-coach-fare-paying-passengers as well has indeed reflected in part an imperfection of competition: they have lost an option they previously enjoyed. At the same time, their choices have been enriched in other ways—by the

proliferation of business class and other such services and frequent flyer benefits.

In part, however, this spillover effect on them reflects the more general characteristic of a market economy that many of the allocative decisions it makes are in effect collective (see my article with William Shew, pp. 229–32): because of economies of scale, what gets produced is dictated by the preferences of the majority—in this case for a lower-cost and quality service than a minority would have preferred.

This deprivation has, however, resulted also from major derelictions by governments. Congestion at major airports at peak travel times (and the consequent inability of passengers to whom time is very valuable to get the delay-free travel they would willingly pay for) obviously means to an economist that the pertinent government authorities have on the one hand failed efficiently to expand airport and air traffic control capacity and, on the other, to price those scarce facilities at their marginal opportunity costs. No wonder there are shortages.

### V. Completing the Balance

This assessment of the “surprises of deregulation” would be grossly distorted if it were not balanced with at least a mention of the respects in which the outcome has either not been surprising at all to its advocates, or the surprises have been happy ones. The last ten years have fully vindicated our expectations that deregulation would bring lower fares, a structure of fares on average in closer conformity with the structure of costs, an increased range of price-quality options, and great improvements in efficiency—made possible by the abandonment of regulatory restrictions and compelled by the greatly increased intensity of competition—all this along with a 35 percent or so decline in accident rates.

### REFERENCES

- Bailey, Elizabeth E., Graham, David R. and Kaplan, Daniel P., *Deregulating the Airlines*, Cambridge: MIT Press, 1985.
- Brenner, Melvin A., “Airline Deregulation—A Case Study in Public Policy Failure,” *Transportation Law Journal*, forthcoming 1988.
- Hurdle, Gloria J. et al. “Concentration, Potential Entry, and Performance in the Airline Industry,” rev., Washington: Antitrust Division, U. S. Department of Justice, December, 1987.
- Kahn, Alfred E., *The Economics of Regulation, II: Institutional Issues*, New York: Wiley & Sons, 1971.
- , (1978a) “Talk to the New York Society of Security Analysts,” New York City, February 2, 1978.
- , (1978b) *Statement on H. R. 11145*, House Public Works and Transportation Committee, Aviation Subcommittee, March 6, 1978.
- , (1978c) “Deregulation of Air Transportation—Getting from Here to There,” in *Regulating Business: The Search for an Optimum*, San Francisco: Institute for Contemporary Studies, 1978.
- , “Applying Economics To An Imperfect World,” *American Economic Review Proceedings* May 1979, 69, 1–13.
- , (1988a) “Deregulatory Schizophrenia,” *California Law Review*, 75, forthcoming 1988.
- , (1988b) “Airline Deregulation—a Mixed Bag, But a Clear Success Nevertheless,” *Transportation Law Journal*, forthcoming 1988.
- and Shew, William B., “Current Issues in Telecommunications Regulation: Pricing,” *Yale Journal on Regulation*, Spring 1987, 4, 191–256.
- Levine, Michael E., “Airline Competition in Deregulated Markets: Theory, Firm Strategy, and Public Policy,” *Yale Journal on Regulation*, Spring 1987, 4, 393–494.
- Maldutis, Julius, *Statement*, Senate Committee on Commerce, Science and Transportation, November 4, 1987.
- Meyer, John R., Oster, Clinton J., and Strong, John S., “Airline Financial Performance since Deregulation” and “The Effect on Travelers: Fares and Service” in John R. Meyer and Clinton V. Oster, Jr. et al, *Deregulation and the Future of Intercity Passenger Travel*, Cambridge: MIT Press, 1987.

Morrison, Steven and Winston, Clifford, *The Economic Effects of Airline Deregulation*, Washington: The Brookings Institution, 1986.

Air Transport Association of America, *Air Transport — Annual Reports of the U.S.*

*Scheduled Airline Industry*, Washington 1975–87.

U. S. General Accounting Office, *Deregulation Increased Competition Is Making Airline More Efficient and Responsive to Consumers*, Washington: November 6, 1985.

# Surprises from Telephone Deregulation and the AT&T Divestiture

By ROBERT W. CRANDALL\*

Undoubtedly, the greatest surprise in telephone industry deregulation has been the absence of deregulation, for the industry continues to be almost as highly regulated today as twenty years ago. Entry has been greatly liberalized in the equipment and most services markets, AT&T has been broken up, but the most important intrastate and interstate telephone services continue to be subject to formal rate regulation. Competitive entry has made this regulation more difficult, not politically less compelling. The major event in the telephone industry has not been deregulation, but divestiture.

In 1984, AT&T was divested of its operating companies as the result of an historic 1982 antitrust decree. In this paper, I summarize some of the early effects of divestiture, including: (i) the virulence of the politics to keep the uneconomic subsidies that invited competitive entry in the first place, (ii) the preliminary evidence that AT&T is "losing by winning" and not vice versa, (iii) the new competition in equipment markets that may turn out to be more important than the recent developments in services competition, (iv) the misplaced concerns about the loss of system efficiency and service quality due to divestiture, and (v) the plight of the divested regional Bell holding companies (RBOCs).

## I. The Politics of Regulation as Cross Subsidization

The success of recent attempts to deregulate transportation may be traced in part to the rather limited extent of cross subsidization caused by the regulation of these markets. As a result, there was limited consumer opposition to deregulation and the

shift to cost-based pricing that would emerge in a competitive market. In the telephone industry, however, the subsidies have been substantial. Long-distance rates have been held artificially high and invariant to differences in traffic density (and therefore marginal costs) by state and federal regulators.

The overpricing of toll services prior to divestiture developed as a convenient means for state regulators to cross subsidize local telephone service. The excess charges on AT&T long-distance calls were transferred to local telephone exchange carriers through a complicated settlements process. As long-distance costs fell, the federal-state regulatory board that controlled regulatory cost allocations decided to allocate an increasing share of the local nontraffic sensitive costs to the interstate long-distance services, thus increasing the degree of cross subsidy.<sup>1</sup>

With the 1984 divestiture, AT&T no longer owns the Bell operating companies through which 80 percent of its calls were originated and terminated. As a result, a system of access fees was established by federal and state regulators to replace the settlements process and to continue the cross subsidies. These access fees are set far above the incremental cost of connecting the calls at each end, thereby permitting the continued flow of revenues from long-distance calls to defray the costs of the local loops.

<sup>1</sup>In this paper, I refer to "cross subsidies" as the assignment of a disproportionate share of the fixed costs of subscriber access lines to long-distance services. This is not to say that this results in long-distance rates above the stand-alone costs of these services in most instances. Rather, it means that the pricing of telephone services is highly inefficient because the joint costs are disproportionately assigned to the service with the higher price elasticity of demand. As bypass develops (see below), this mispricing will indeed be shown to be a subsidy by even the most stringent definition.

\*Senior Fellow, The Brookings Institution, 1775 Mass. Ave., NW, Washington, D.C. 20036.

The regulatory overpricing of access charges for long-distance calls inevitably will lead AT&T and the other long-distance carriers and their customers to seek alternatives to the local telephone companies for connecting their calls. Unfortunately, this "bypass" threat has developed slowly, thereby allowing politicians to continue to clamor for the cross subsidization of local service from long-distance access charges. To its credit, the FCC has recognized the economic welfare losses inherent in this mispricing of telephone services, but it has been thwarted by pressures from Congress in its attempts to substitute a more efficient system of fixed monthly subscriber line charges to defray the costs of the telephone plant that connects each subscriber to the local switch.

After two years of phasing in the monthly subscriber line charges as a substitute for interstate long-distance access charges in the face of immense congressional opposition, the FCC has achieved only a modest repricing of telephone services through these subscriber line charges. In 1986, subscriber line charges accounted for a mere 3.7 percent of local telephone company revenues while access charges remained at 28 percent of revenues, the same percentage as in 1984 (the first year of divestiture). (See FCC, *Statistics of Communications Common Carriers*, 1984; 1986.) Even at the current (1987) residential subscriber line charge, only about one-third of the nontraffic sensitive costs assigned to interstate switched services are defrayed by direct monthly subscriber charges. Thus, a very large share of the pre-divestiture cross subsidies continue to be extracted from long-distance services.

The results of the attempts to reprice telephone rates can be seen in recent telephone rate trends. Local access/exchange rates have risen in real terms while long-distance rates have fallen rather dramatically. The increase in local rates began in 1980, however, before the imposition of subscriber line charges as state commissions began to allow telephone companies to recoup costs not recovered during the inflationary 1970's. Real local rates have continued to rise at about 5 percent per annum since divestiture with the

repricing of services caused by the subscriber line charges, while real interstate long-distance rates have declined by about 8 percent per year since divestiture. (See BLS, *Consumer Price Indexes for Telephone Service*.) State regulators have been much more reluctant to allow long-distance rates to fall in intrastate markets, preferring to continue the cross subsidization of local service. As a result, intrastate toll rates were unchanged in real terms between 1982 and 1986.

Politicians have decried the limited repricing that has occurred, seeing it as a threat to universal service. But the evidence does not support such a concern. First, the price elasticity of demand for local access services is extremely low—about  $-0.05$  to  $-0.2$ . (See Lester Taylor, 1980, p. 80.) Second, there is no evidence that the share of low-income households subscribing to local telephone service has declined. In fact, perhaps because of "lifeline" rates in some states, telephone subscription appears to have risen among low-income households. In November 1986, 92.4 percent of all households had telephone service as compared with 91.4 percent on the eve of divestiture. (See U.S. Bureau of the Census, *Current Population Survey*.) At household incomes below \$7,500 per year, telephone penetration actually increased during this period.

## II. AT&T vs. the RBOCs—Who is Winning?

One view expressed by some students of the industry after the AT&T divestiture was that AT&T won by losing the antitrust battle since it was able to spin off its troglodyte operating companies and keep its high-tech equipment and interstate services operations. (See Paul MacAvoy and Kenneth Robinson, 1983.) The early evidence on the returns to stockholders suggests that this view may have been oversold. The cumulative return from holding the Regional Bell Holding Company equities has been about double the rate from holding AT&T common (data from CRSP tapes). Given the lower systematic stock market risk of holding RBOC equities, this difference cannot be ascribed to a  $\beta$  effect in a rising market. Hence, one must conclude

that the RBOCs have done substantially better than AT&T relative to early 1984 expectations.

### III. Competition in Equipment Markets

Perhaps the most startling result of the AT&T divestiture has been the sharp increase in competition in the switching and transmission equipment markets. AT&T's Western Electric was once assured of the market for most of the requirements of its Bell operating company affiliates. Since divestiture, however, AT&T has lost more than one-third of this business to competitors (see my forthcoming paper). Surprisingly, very little of this loss is in large switches or PBXs, where AT&T has been able to meet the surge in competition reasonably well. Rather, the losses appear to be in the transmission equipment, wire, and various other products that go into the telephone company plant.

It is too early to discern any effect of divestiture on technological progress and the prices of telephone equipment, but the early indications are favorable. Switch prices per line are falling at about 11 percent per year (see U.S. Department of Justice, 1987, p. 14.11), and the introduction of new features appears to be accelerating. There is some evidence accumulating that telecommunications switching technology historically evidenced less progress than the comparable electronic computer technology. By increasing competition in equipment supply, the divestiture may have spurred a closing of this gap (see Kenneth Flamm, forthcoming).

### IV. Quality and Output

One of the principal concerns over divestiture was that a fragmented telephone system—with separate operating companies, long-distance companies, and equipment companies—would provide lower-quality service and reduced output. In the first year of divestiture, there were substantial problems in obtaining prompt installation or repair service. But since 1984, there have been so few problems that it is difficult to find any state commission that is concerned with

transmission quality, dial tone delay, or circuit blockage.

There is, however, a mild erosion in the *quantity* of telephone company service since divestiture because of competition. An increasing share of telephone service has been moving away from the regulated telephone sector since the 1970's when the FCC began to allow competitive entry into interstate services, cellular (mobile) services, and terminal equipment. In addition, terminal equipment and inside wiring has been de-tariffed. This erosion of the common-carrier share of total telephone services has continued with divestiture, but it does not appear to have accelerated. At present, roughly half of all new telephone investment is being undertaken by entities other than the local or long-distance telephone companies.<sup>2</sup> With the Bell operating companies limited to local service, this shift of telephone services from regulated to unregulated carriers will clearly continue and may well accelerate.

### V. The Plight of the RBOCs

Although the divested regional Bell operating companies have outperformed AT&T in the equity markets since 1984, their long-run prospects are quite uncertain. Under the decree, these companies are forbidden to engage in equipment manufacturing, inter-LATA long-distance service, or information services. In the first triennial review of this provision of the 1982 antitrust decree, Judge Greene ruled that until there is substantial competition in the provision of local connections to interstate telephone service, these restrictions would not be lifted (see *U.S. v. Western Electric et al.*, 1987).

The theory of the 1982 AT&T decree is that the integration of long-distance services or equipment manufacture with the "bottleneck" local access service provided the local Bell Telephone companies with the ability to frustrate competitive entry into these markets. Unfortunately, limiting the

<sup>2</sup>My estimate from Department of Commerce data.

bottleneck monopolist to the local market, in which rates are kept artificially low by regulators, creates a rather paradoxical problem. The overpricing of interstate access fees inevitably creates the incentive for large customers or long-distance competitors to find alternative technologies to connect these large customers to the interexchange switch. On the other hand, because local telephone rates are set at levels that are less than costs, there is little incentive for new entry into the market for supplying local service to dispersed customers with a low ratio of long-distance to total calls. Over time, the divested RBOCs will face the loss of their large highly profitable customers, who in turn, will avoid having to contribute to the subsidization of the smaller customers. Thus, like the railroads in earlier decades, the Bell operating companies may find themselves left with a monopoly over the low-margin business and with little of the high-margin business. But under the reasoning of the decree, their nonremunerative monopoly is the primary reason for keeping the RBOCs out of all other lines of business in the telephone industry.

## VI. Prospects for Deregulation

The message of this paper is that despite the increase in competition in long-distance, enhanced services, and equipment markets, the distortions caused by regulation of the telephone industry continue. The political justification for continued regulation (and its attendant distortions) is the monopoly bottleneck of the local telephone exchange and the need to protect local service. As we have seen, the concern over universal service has been greatly exaggerated. The bottleneck monopoly issue is more complex, however, and cannot be discussed in detail in this paper. It is sufficient to point out that competition for dispersed residential and smaller commercial customers is not likely to be intense unless cellular systems decline sharply in cost and the cross subsidies that support local service are attenuated.

One should not conclude, however, that the political demand for cross subsidies will

stifle all attempts to deregulate the telephone industry. The FCC is now examining alternatives to traditional rate-of-return regulation for AT&T. Complete deregulation of interstate services is not out of the question given the extensive investment by new competitors (OCCs) in fiber-optic capacity. This deregulation would probably not include the elimination of access charges that now distort relative prices. In a few years, however, bypass and the gradual increase in subscriber line charges may erode these regulatory cross subsidies substantially.

There is also a substantial movement for "regulatory flexibility" or even deregulation of local and intrastate services in some states. The prospects for deregulation in such states as California or Massachusetts, however, is quite limited. The pressure for regulatory cross subsidies in these states is not likely to subside until everyone is carrying a telephone in his or her pocket.

## REFERENCES

- Crandall, Robert W.**, "Structural Separations and the Role of the Telephone Operating Companies," in Robert W. Crandall and Kenneth Flamm, eds., *Changing the Rules: Technological Change, International Competition, and Regulation in Communications*, Washington: The Brookings Institution, forthcoming.
- Flamm, Kenneth**, "Economic Dimensions of Technological Advance in Communications: A Comparison with Computers," in R. W. Crandall and K. Flamm, eds., *Changing the Rules...*, Washington: The Brookings Institution, forthcoming.
- MacAvoy, Paul W. and Robinson, Kenneth**, "Winning by Losing: The AT&T Settlement and Its Impact on Telecommunications," *Yale Journal on Regulation*, No. 1, 1983, 1, 1-42.
- Taylor, Lester D.**, *Telecommunications Demand, A Survey and Critique*, Cambridge: Ballinger, 1980.
- Federal Communications Commission**, *Statistics of Communications Common Carriers*, 1984 and 1986.
- U.S. Department of Commerce**, Bureau of the



Census, *Current Population Survey*, various issues.

U.S. Department of Justice, *The Geodesic Network: 1987 Report on Competition in the Telephone Industry*, Washington, 1987.

U.S. Department of Labor, Bureau of Labor

Statistics, *Consumer Price Indexes for Telephone Service*, various issues.

*U.S. v. Western Electric, et al.*, Civil Action 82-0192, U.S. District Court of the District of Columbia, Order, September 10, 1987.

# Interaction of Financial and Regulatory Innovation

By EDWARD J. KANE\*

What I find surprising about the phenomenon of "financial deregulation" is economists' insistence on thinking about regulatory adjustments that affect financial firms as exogenous disturbances to a general economic equilibrium. Far from being a politically self-contained disturbance to financial markets, "deregulation" is an endogenous response by regulators to changes in the economic constraints that financial markets impose upon them.

My perspective on financial and regulatory innovation may be grasped by visualizing the front window of a *large* financial-services firm. In this window are four signs. Three of the signs constitute electronic displays. The messages on these three signs as well as the equipment used to display them are continually updated by the firm's employees. The three signs display respectively the following information:

1) The *product lines* the firm offers: different types of deposit or investment accounts, credit arrangements, and other customer services; 2) The *prices* the firm currently attaches to each type of product; 3) The *name, office locations, and organizational form* of the institution itself.

What about the fourth sign? This one is painted permanently on the window in gold letters. It says that the debts of this institution are *guaranteed in full* by either its home or host government because the firm is too large for affected politicians to allow it to fail.

This image hints at two points. First, the permanence of the information conveyed by the fourth sign and the slowness with which politicians and bureaucrats adjust their monitoring of institutions' risk-taking activ-

ity to changing opportunities for taking risk help to explain the impermanence or volatility of the information displayed on the other three. Underpriced and insensitively monitored government guarantees cushion the penalties from failure that ordinarily constrain innovative behavior. Second, government guarantees and supporting regulatory activity are only part of the story. The other major forces are volatility in financial firms' macroeconomic and microeconomic environments, particularly the rapid technological change symbolized by the electronic signs whose form and content the firm's managers directly control.

Financial theory holds that financial firms exist to reconcile in an economical fashion the funding needs of entities that want to spend more than their income with the desire for credit-enhanced savings vehicles on the part of entities that want to accrue a surplus. Conventional theory portrays society's savings propensities, the productivity of real capital, fiscal and monetary policy, and the technology of information processing and financial transacting as determining both the prices at which a financial-services firm could afford to offer untaxed and unsubsidized financial products and the essential economic functions it seeks to perform. My research (1984; 1987) takes these elements of the problem as given. It stresses that, overlaying the pattern of financial opportunities, regulatory competition helps to shape the formal organization of the firm. By "organization," I mean the details of a financial intermediary's corporate structure, the locations and processes it uses to produce and distribute financial services, and the names and contractual details of the financial instruments that constitute its product line. My analysis stresses further that regulatory burdens and subsidies and regulatee adaptation to them simultaneously determine each other.

\*Ohio State University, Columbus, OH 43210.

### I. Concepts With Which to Contemplate the Changing Landscape of Finance

To encapsule the expanding range of activities being undertaken by contemporary financial organizations, we need generic definitions of a financial-services firm (FSF) and of financial regulation. An FSF produces informational and transactional products for a base of customers with whom it establishes client relationships. To deliver any financial service, an FSF must exchange information with its customers. This definition clarifies why technologies of communication, information storage, and data processing stand in the forefront of modern financial activities. To exchange information requires information media. These media specifically connect the customer with the FSF product that is desired. Information media run a gamut from person-to-person contacts, paper evidences, and telephonic messages to magnetic coding, keyboard-actuated video displays, and sophisticated kinds of electronic imagery. The increasing use of robotic mechanisms to exchange information and effect transactions in the financial industry suggests the whimsical possibility of a bank robber sending out a robot to hold up an automated teller machine for him.

The perceived quality of an FSF's products increases with the confidence and convenience its customers attach to them. Economic efficiency is served by an FSF's arranging to produce financial services jointly with an external supplier of regulatory services. Third-party monitoring, disciplinary, certification, and guarantee services aim either at promoting customers' confidence in an FSF's ability to perform or at coordinating competitor activity to enhance the transactional convenience of an FSF's products.

Also useful in understanding financial change are Schumpeterian distinctions between inventors and entrepreneurs and invention and innovation. An invention is an unfolding technological opportunity: the discovery either of a way to do something that has never been done before or of a better way to perform a longstanding function. In-

novation is the act of applying an invention: putting an inventive idea into profitable operation. Typically, delays occur between the appearance of an invention, the discovery of its commercial potential by an entrepreneur, and its embodiment in a concrete innovation. Delays between invention and innovation may be termed discovery and execution lags.

The financial-intermediary business used to be a comfortable, largely noninnovative one. Managers of financial institutions operated within a relatively fixed environment and at a fairly leisurely pace. Now they have to develop reliable information more quickly, to make quicker decisions, to watch their competitors more closely, and to look constantly for new ways to serve customers and to organize their firm's affairs. They need to possess more knowledge, more imagination, and fancier equipment. They have to identify new powers that could make their business more profitable and transform the names of their firms in ways that can communicate to customers and staff the expanded geographic and functional reach to which the firm now aspires. Finally, they must develop the political savvy to persuade politicians and regulatory authorities to let them move into new turf.

Parallel comments apply to financial regulators. Within and across countries, existing patterns of exclusionary regulation are crumbling. In attempting to rebuild their domains, authorities are finding regulatory positions in financial markets hard to sustain. To analyze financial and regulatory innovations in parallel fashion, it is instructive to conceive of regulatory entities as multi-product firms and to explore the evolution of entry and exit costs in the market for financial regulatory services in which regulators operate.

A market may be defined as a collection of persons carrying on extensive and at least partly voluntary transactions in a specific good or service. William Baumol, John Panzar, and Robert Willig (1986) define an individual market as "perfectly contestable" when the costs of entering or exiting that market are zero. In any such market, the

threat of hit-and-run entry by outside potential competitors holds the profit margin sought by incumbent firms to competitive levels, irrespective of the number of incumbent competitors or of how concentrated industry output might happen to be.

Regulatory services are typically delivered in the context of an ongoing client relationship rather than sold on a transaction-by-transaction basis. Regulatees contract for a vector of contingent services without necessarily haggling specifically over the prices of individual services. Nevertheless, in an important sense, some of a regulator's clients are always shopping for a better regulatory deal. Whenever a regulatee fears that its traditional regulator's prices stack up poorly to the competition, it must study its options for switching some or all of its business to a new servicer. In practice, such study puts pressure on an FSF's current principal regulatory supplier to develop a more favorable set of prices or an improved level of service.

Markets for regulatory services are demonstrably *not* perfectly contestable. Significant exit costs exist. The incompleteness of public accounting systems allows government regulators to conceal large implicit losses in the short run. Economically unprofitable government regulators can, when pressed for survival, deliver subsidies that hold old clients and attract new ones. Even in the absence of subsidies, regulatees that try to switch to a new supplier of regulatory services often incur substantial transition or switching costs.

Analysis of financial change must focus on the capacity of different regulatees and regulators to adapt to exogenous and endogenous decreases in the costs of entering and exiting different financial product markets. Product-line and geographic-market expansion by suppliers of financial regulatory services follows and supports rivalry between client FSFs within and across countries, regions, and various kinds of administrative boundaries. Without denying bureaucratic aspects of regulatory behavior (William Niskanen, 1971), my explanation of regulatory innovation focuses on regulators' efforts to improve their market posi-

tion. A convenient way to model these efforts is to posit that, subject to defective financial reporting and profitability constraints that permit government entities to conceal implicit costs and to accelerate implicit revenue, regulators seek to extend or to defend their share of the market for regulatory services in the face of exogenous and endogenous disturbances in their economic environments.

## II. Globalization and Product-Line Fusion as Structural Arbitrage

These definitions and distinctions clarify why financial regulation is endogenous. To call the global integration of financial markets the result of either exogenous deregulation or exogenous technological change is to miss the interactive nature of the adjustments that are taking place. On one side of the process, regulatees are changing their product lines, office locations, production and delivery processes, and organizational forms both to take advantage of emerging technological opportunities and to lighten net tax and regulatory burdens. Calling this adaptation *structural arbitrage* underscores the notion that timely changes in the structure of a firm's operations can create profits just as surely as the activity of buying something cheap and selling it dear. From regulators' perspective, structural arbitrage creates costs and benefits for their enterprises that change their ideas of what constitute optimally designed national or subnational tax codes and regulatory arrangements. Recognizing this leads us to see realignments in applicable tax and regulatory frameworks as largely reactive acts of competitive reregulation.

Regulatory interference imposes entry restrictions and corresponding avoidance costs on expanding firms. But, in a free society in which multiple legislatures and regulatory agencies compete for regulatees, tax receipts, and budget funds, authorities can only induce great or long-lasting divergences between the actual and the cost-minimizing global financial-market structure when the costs of structural arbitrage are high. As

these costs fall toward zero, efficient patterns of resource allocation displace inefficient ones.

During the last twenty years, technological advances in information processing, robotics, and telecommunications have regularly lowered the distance-related entry and enterprise-coordination costs confronting firms that operate in diverse and far-flung financial-services markets. During the same interval, increasing volumes of multinational production and world trade combined with shifting patterns of balance-of-payments surpluses and deficits to increase greatly the rewards that large FSFs could expect to earn from adapting their operations to span and integrate financial markets multinationally.

Around the world, FSFs have been circumventing regulatory barriers to entering each other's traditional lines of business and geographic markets and transforming their front offices into partially robotized multi-product bazaars and their back offices into electronic transactions and communications centers. Advances in electronic and financial-contracting technology have played a major role in these developments by creating unregulated or less-regulated "loophole" substitutes for tightly regulated traditional products and ways of doing business.

Transformation of front- and back-office production and delivery systems is exemplified by the expanding transactional capabilities of successive generations of automated teller machines (ATMs). Today, ATMs can be supported by partially automated loan-application processes and on-line back-office computer systems employing credit-decision software and linked through multinational networks and interchanges. Users of an ATM network can transact (often via satellite transponders) with their local deposit institution from a substantial subset of roughly 175,000 ATMs.<sup>1</sup> In effect, a shared

ATM is not just a robot substitute for a teller, but a loophole substitute for a limited-service branch office at the ATM's location and for offers of higher explicit rates of interest.

Entry into nontraditional lines of business is exemplified by development of loophole financial instruments and loophole forms of corporate organization that fuse what used to be institutionally and regulatorily disparate product lines. Among the sharpest symbols of this fusion are securitized loans and deposits and the diversified financial-services holding company.

Securitized loans are a fund-raising technique that substitutes for asset sales or loan participations. A loan is "securitized" when it is packaged into an intelligible (i.e., rateable) collateral pool, whose cash flows back an issue of securities sold by the loan's originator or servicer. Although securitization has been most widely used to borrow against the collateral of consumer mortgages and automobile loans, pools of commercial mortgages, equipment leases, junk bonds, insurance-premium cash flows, and even poorly performing loans have also been securitized. Securitization can be used to unbundle a traditional lender's origination, servicing, credit-rating, risk-bearing, and financing functions. The borrower's obligations and rights under a loan contract can be unbundled too, by pledging various time-dated cash flows to different "strips" of a collateralized securities package as in a collateralized mortgage obligation. One U.S. lender (Perpetual Mortgage Co.) has gone so far as to make its borrowers sign a series of separate notes at the loan closing.

A deposit is securitized when the return it promises is linked to an index measuring the performance of a specific portfolio of risky

<sup>1</sup>According to Jeffrey Kutler (1987) seven principal ATM networks exist: Express Cash (20 countries, 23,795 machines, and 25 million cardholders); Visa (21 countries, 19,203 machines, and 150 million cardholders);

Cirrus (U.S. and Canada, 17,200 machines, and 62 million cardholders); Plus (U.S. and Canada, 14,000 machines, and 70 million cardholders); Master Teller (6 countries, 9,462 machines, and 130 million cardholders); ADP Exchange (U.S. and Canada, 4,500 machines, and 8 million cardholders); Eurocheque (Europe and Middle East, 2,500 machines, and 31 million cardholders).

assets. First offered by Chase Manhattan Bank in March 1987, indexed deposits cross the downside protection of a government-insured deposit with the upside potential of a securities investment (typically the Standard & Poor's 500 index). Whereas securitized loans raise funds by selling deposit-institution debt that is backed by *concrete* assets into wholesale capital markets, indexed-deposits sell limited participations in a *hypothetical* short or long position in a securities portfolio both into wholesale markets through securities firms and to a deposit institution's traditional customer base. Issuers of indexed deposits hedge the product by purchasing options and future contracts on the underlying portfolio. Although volatile world stock markets should offer a near-ideal environment in which to market what are principal-protected "bear" and "bull" bets on the course of future stock prices, volatility increases the issuer's cost of hedging. These costs limit the proportion of an index's appreciation or depreciation an issuer can afford to pay out to depositors.

Changes in corporate form can change restrictions on product line and office locations and even the particular government or self-regulatory agencies that write, administer, or enforce these restrictions. In selecting a particular set of structural options, an FSF chooses what we may presume to be an optimal "tax and regulatory microclimate." The dimensions of this climate include: charter type (for example, bank vs. security firm), chartering agency (typically, a national or subnational governmental entity), ownership structure (mutual vs. stock; direct vs. indirect ownership) and cross-organization control linkages (branch office or other forms of corporate presence vs. subsidiaries vs. holding-company affiliation).

### III. The Regulatory Dialectic: Economic Constraints on Regulators

Regulation endeavors to set unwelcome rules on someone else's behavior. These rules seek either to forbid or to compel particular kinds of behavior on the part of a designated set of regulatees. Regulatees (including most children) learn a series of what we may call

circumvention or avoidance behaviors. Regulation and avoidance are as hard to separate as Siamese twins. Rules and loopholes coexist in every legal text and in every regulatory system.

The tandem nature of regulation and avoidance is featured in a conceptual framework designed to dramatize the process of financial and regulatory innovation: the regulatory dialectic. *Dialectic* is philosopher's shorthand for a process driven by tension between a succession of paired opposites. Dialectical outcomes are governed by the push and pull of opposing forces. Movement comes from ongoing conflict and conflict resolution between opposing ideas and the logical, physical, political, or economic forces associated with them.

The philosopher Hegel named the opposing ideas the *thesis* and *antithesis*, and called the idea that develops to resolve their conflict the *synthesis*. The restlessness featured in dialectical thinking comes from the notion that each synthesis becomes a thesis in a new dialectic. This view sees the evolution of thinking on any issue as driven by a *three-stage cycle* in which every idea first calls forth opposition from a conflicting idea. Then, whatever idea resolves a given conflict is immediately confronted with a contradictory proposition so that the cycle is perpetually renewed.

I can illustrate the same point by drawing an analogy with playground games such as chase or tag. When these games are pursued by serious players, opponents work hard to stay out of each other's reach and to unbalance the inherently temporary victory won at each change of initiative.

To apply dialectical thinking to the regulatory scene, the thesis and antithesis may be identified with *regulation* and *avoidance*, and the third stage renamed *reregulation*. In the dialectical view, all regulation becomes *reregulation*, whose shape is determined by the precise history of prior reregulatory problems. Over any finite time interval, two alternative sequences may be distinguished, depending on whether regulators or regulatees are viewed as kicking off the adaptive process: 1) Regulation-avoidance-reregulation and 2) Avoidance-reregulation-avoidance.

The regulatory dialectic portrays regulation as one side of a game of strategy with sequential moves. In this game, the players on the various sides react to one another in creative ways. Exploiting the Schumpeterian distinction between invention and innovation, we can enrich the model by positing differences in the speed with which different types of players characteristically respond to their opponents' moves.

Hypothesize first that innovation discovery and execution lags are typically shorter for regulatees than for regulators. Also hypothesize that discovery and execution lags are shorter for less-regulated competitors than for a specific regulation's targeted set of regulated players and that regulatory lags are shorter for industry self-regulators than for government bureaucrats. To justify these hypotheses, let us appeal to differences in relevant information costs, differences in the extent of managerial commitment to the goals of regulation, and differences in the extent to which principal-agent conflicts can be resolved in government and private enterprises.

#### IV. Financial Instability as the Cost of Inefficient Financial Regulation

The strength of a dialectical vision is the evolutionary perspective it gives us for confronting and interpreting change. The regulatory dialectic has two policy implications. First, in the face of exogenous changes in technology and economic volatility, rooting policies in concepts of stationary equilibrium is unreliable. Even if (as U.S. authorities desperately wish) a global cartel in financial regulatory services were successfully to be negotiated, the cartel would contain the seeds of its own future destruction. Second, the problems being experienced by any set of regulatees and regulators is rooted in the detailed history of their prior conflict. For this reason, would-be regulatory reformers need to look beyond immediate problems to assess the long-run consequences of the policies they wish to install.

The regulatory dialectic emphasizes that, in the long run (by which is meant a period long enough that adjustment and informa-

tion-acquisition costs become irrelevant), survivable patterns of regulation must be economically efficient ones. But even though the invisible hand eventually punishes over- and under-regulators alike, in real time the process can produce considerable turmoil. The sequential search for efficiency can take a long time to unfold and can impose substantial pain of FSFs, their customers, and the general taxpayer.

From the point of view of their regulatees, revenue losses imposed by regulators' explicit charges and various operational constraints reduce the net value of the regulatory services received. We may define the balance between the costs and benefits that a given regulator succeeds in imposing on its regulatees as their net regulatory "burden" (or subsidy). The regulatory dialectic posits a dynamic adjustment process that in the long run enforces a "law of one regulatory burden." Precisely because inefficient patterns of regulation impose excessively burdensome costs either on regulatees, their customers, or the general taxpayer, the burdened parties must be expected *sooner or later* to develop avoidance strategies by which to throw these burdens aside. However, the more effectively a given set of regulators can hide the financial burdens from those who ultimately bear them, the longer it will take for effective avoidance strategies to come into play.

The variable nature of burden discovery and avoidance lags clarifies both what can go wrong in regulatory competition and why it is nevertheless a mistake to view rivalry among alternative regulators for clients and budgets merely as wasteful duplication. A monopoly supplier or regulatory cartel would tend in the short run to overregulation. When burden-bearers and elected politicians are well-informed, overlaps in regulatory missions across different regulatory entities promote short- and long-run efficiency in the production and delivery of regulatory services, much as duplication of service functions across private institutions promotes efficiency in the provision of financial services. However, when a regulator and its clients can exploit and perpetuate impediments in burden-bearers' access to the infor-

mation needed to judge the regulator's performance, this competition can *temporarily* promote inefficiency instead. In the short run, inappropriately monitored regulators can deliver unintended and economically inappropriate subsidies. Only when the burden-bearers find a way to enforce their interest in preventing subsidies from being hidden, can we say that interregulator rivalry protects borrowers, depositors, and investors from the short-run as well as the long-run dangers of underregulation.

Some U.S. authorities are currently working very hard to prolong underpriced and misadministered deposit-insurance guarantees, selected restrictions on deposit-institution interest rates and product lines, and vestigial prohibitions against interstate banking. These efforts to prolong inefficient patterns of financial regulation help to conceal subsidies to risk bearing that increase economic volatility and threaten to disrupt world financial stability in the short run.

## REFERENCES

- Baumol, William, Panzar, John C. and Willig, Robert, "On the Theory of Contestable Markets" in G. F. Mattavson and Joseph E. Stiglitz, eds., *New Developments in the Theory of Industrial Structure*, Cambridge: MIT Press, 1986.
- Kane, Edward, J., "Technological and Regulatory Forces in the Developing Fusion of Financial-Services Competition," *Journal of Finance*, July 1984, 39, 759-72.
- \_\_\_\_\_, "How Market Forces Influence the Structure of Financial Regulation," mimeo., Ohio State University College of Business, October 1987.
- Kutler, Jeffrey, "Visa Reports Its ATM Network Is Largest Bank-Owned System," *The American Banker*, August 14, 1987, 152, pp. 1 and 15.
- Niskanen, William, *Bureaucracy and Representative Government*, Chicago: Aldine, 1971.



**THE WIDESPREAD DEPRESSION OVERSEAS:  
AMERICAN AND PACIFIC INFLUENCES<sup>†</sup>**

**On Macroeconomic Implications of Price Setting  
in the Open Economy**

By JEAN-PAUL FITOUSSI AND JACQUES LE CACHEUX\*

Ever since its inception, the world system of floating exchanges rates has been characterized by large and persistent movements in currency values, both nominal and real, with apparently no tendency for purchasing power parity to assert itself either in the aggregate or on a product basis; in particular, there is growing evidence that the "law of one price" does not seem to hold even for tradables and that price-cost markups display large and persistent fluctuations. In the 1980's, these movements have been accompanied by worldwide high real interest rates and ample, long-lasting differentials, as well as persistent trade and current account imbalances on a grand scale for many countries. In addition, there has been widespread concern that expansionary policies in one large country—namely the United States—may have had adverse effects on others, European countries in particular.

Such observations and opinions are apparently difficult to reconcile with standard open-economy theory, as synthesized, for instance, by Rudiger Dornbusch (1980), or with more "classical" models of the recent vintage. In a recent monograph, Fitoussi and Edmund Phelps (1988) showed that when firms are assumed to set prices in a way consistent with "customer market" behavior, international transmission is effected also via

supply or markup responses, even when goods are traded and goods markets have a competitive structure. But, as shown by the revived interest in imperfect competition, price setting can arise in a large number of market environments. While the macroeconomic consequences of price setting in the closed economy have been the object of numerous recent contributions, applications to open economies are still in their infancy. The purpose of this paper is to sketch some (admittedly rather crude) working hypotheses and explore some of their macroeconomic implications that appear to be specific to open economies.

**I. Alternative Foundations  
of Price-Setting Behavior**

That price determination outside the very restrictive set of assumptions of Walrasian auction-market pricing is likely to differ markedly from the result of simply equating supply and demand has long been a major concern of economic analysis. The conditions for price-setting behavior are always such that the demand curve perceived by the firm is not infinitely elastic with respect to the price it charges for its product. Such characteristics used to be thought of as narrowly confined to cases of imperfectly competitive market structures, in which there will be strategic interactions between a small number of firms. However, when the concept of market imperfections is enlarged to include informational imperfections and consumer search, even firms operating in a fairly competitive environment may benefit from some—transient or long lasting—market power (Phelps and Sidney Winter, 1970).

<sup>†</sup>*Discussants:* Matthew B. Canzoneri, Georgetown University; John B. Taylor, Stanford University; Rudiger Dornbusch, MIT.

\*Professor of Economics, Institut d'Etudes Politiques, Paris, and Director, Research Department, O.F.C.E., 69 quai d'Orsay, 75007 Paris; and Deputy-director, Research Department, O.F.C.E., Paris, and Maître de conférences, I.E.P., Paris, respectively.

With consumer search, price-setting decisions by individual firms will be the result of complex conjectures, due to the existence of strategic interactions both with competing firms *and* with consumers. The outcome will therefore depend to a large extent on the precise specification of exchange and information technologies (see, for example, Joseph Stiglitz, 1987); in many cases, though, perceived demand curves of individual firms will be kinked at the going price, which entails major departures from the usual assumptions regarding price adjustment in competitive environments. Specifically, individual prices will not generally be competed down to marginal cost, that is, there will be positive markups in equilibrium; also, there will exist an equilibrium price distribution and individual prices will not be adjusted in response to shocks that move the equilibrium within a well-defined range. Consequently, individual prices are bound to change infrequently and to be determined by pricing conventions.

Infrequent price revisions also arise as a result of firms' optimizing behavior in the case when price changes involve "menu costs" (see, for example, Olivier Blanchard, 1983; George Akerlof and Janet Yellen, 1985). Under a wide variety of reasonable assumptions with regard to exogenous monetary processes, the existence of (even small) menu costs will also induce staggered price-setting by individual firms and some degree of stickiness in the aggregate price level.

## II. Staggered Price Setting in the Open Economy

In the field of international economics, models of imperfect competition have recently been proposed to explain the response to exchange rate changes (see, for example, Elhanan Helpman and Paul Krugman, 1986; Krugman, 1986; Dornbusch, 1987; Le Cacheux and François Lecoq, 1987). These analyses suggest that, with imperfect competition, firms may be able to charge a different price for their product on each specific market, (i.e., to "price to market"). They show that, in such cases, the elasticity

of import prices—expressed in domestic currency units—with respect to the nominal exchange rate will be less than one and that there may be persistence in real exchange movements, as well as in trade balance adjustments, in response to exchange rate changes (Krugman and Richard Baldwin, 1987). However, existing analyses along these lines have been primarily confined to partial-equilibrium investigations of the consequences of exogenous changes in exchange rates on a single country's imports, with the cross-country effects left implicit.

By now, it is a common observation that nominal exchange rates move a lot in flexible rate systems; it is also generally agreed that day-to-day variations are determined in the financial markets in response to "news." Clearly, too, goods prices, whether in the aggregate or individually, are usually not changed quite as frequently: staggered price setting would therefore appear to be an appropriate working hypothesis for the analysis of exchange rate influences on goods prices, in addition to being consistent with a large number of the stories that can be told at the micro level, since the various foundations recounted in Section I are in no way mutually exclusive.

To make the point in the simplest possible way, let us consider a two-country world in which all goods are traded, with goods markets being somehow imperfectly competitive and characterized by consumer search. Rather than explicitly deriving optimal pricing rules from first principles, it will be convenient to appeal to simple, generic pricing conventions that are broadly consistent with the underlying microeconomic assumptions discussed above.<sup>1</sup> Firms are assumed to be based in one country and to sell their products on both markets, all goods being at least imperfect substitutes in demand. There is a large, but finite, number of firms; con-

<sup>1</sup>As is well known, this pragmatic analytical procedure gives rise to behavioral relations that may not be policy-invariant, but doing otherwise would require a complete specification of search technologies and individual firms' strategies. It should be clear, however, that pricing rules are likely to depend on the exchange regime.

sumers search in their home market, but cannot engage in search in the foreign market, so that firms may discriminate among the markets they provide, hence generally "pricing to market." Each firm's perceived demand curve on every market is downward sloping and may be expressed as

$$(1) \quad N_{j,t} = A(P_{j,t}/Q_t)^{-\beta}$$

where  $N_{j,t} = Z_{j,t}/Z_t$  is firm  $j$ 's market share at time  $t$ , (i.e. its individual demand divided by total demand);  $P_{j,t}$  is firm  $j$ 's posted price and  $Q_t$  is an index of market prices.  $\beta$  is an elasticity parameter ( $\beta > 0$ ) depending on consumer preferences and search technologies.

Due to "menu costs" of changing prices and/or to the demand curve being kinked at the going price, pricing decisions will be revised only infrequently and in response to large enough changes in market conditions. In the simplest case, when price tags are posted for a fixed length of time—assumed to be longer than the periodicity of exchange rate changes—and individual prices may be revised at the beginning of each successive period, we get a Taylor-like, staggered price-setting behavior. Individual firms set their price on each market so as to maximize expected future profits; which, in this case, is equivalent to maximizing the current-period expected flow of profit, defined in a standard fashion as

$$(2) \quad \Pi_{j,t} = Z_{j,t}(P_{j,t} - C_{j,t})$$

with  $C_{j,t}$ , the unit cost of production expressed in the same units as the price. In order to focus on the consequences of staggered price setting in the open economy, we assume that production technologies are characterized by constant marginal costs and that domestic costs do not vary.<sup>2</sup> This leaves exchange rate changes as the only possible source of relative cost variations. The pricing

rule of any single firm on each market will therefore have the following, generic form

$$(3) \quad P_{j,t} = F_j(E_t Z_t, E_t Q_t)$$

where  $E_t$  is the expectational operator, conditional on information available at time  $t$ ;  $Z_t$  and  $Q_t$  are averages over the period when the price is fixed. With asynchronous staggered price setting, the aggregate price index  $Q_t$  will, at any time, depend on past and current price decisions.

Since there are both foreign and domestic firms in the market, individual pricing decisions will be influenced by the average exchange rate that is expected to prevail during the period between successive price revisions; and so will the aggregate price index. When individual demand curves are downward sloping and firms face a tradeoff between present profits and market shares, the elasticity of individual prices with respect to the expected exchange rate will be less than one and will depend on initial market shares, provided competition is of the Nash variety. Solving this pricing problem and aggregating will lead to an aggregate price equation—and an aggregate supply equation—which displays some degree of inertia, with both backward-looking and forward-looking characteristics, even assuming perfect foresight (see Taylor).

Ignoring the possible kink in demand curves, which would give rise to asymmetric pricing rules, this may be expressed in a linearized form as follows

$$(4) \quad q_t = \alpha \cdot q_{t-1} + \varepsilon \cdot E_t z_t + \eta \cdot E_t x_t, \quad 0 < \eta < 1$$

where constant terms have been omitted and lowercase letters stand for percent deviations in the variables,  $X_t$  being the nominal exchange rate, defined as the home-currency price of foreign currency.

When the firm's current price decision has consequences that extend beyond the period for which the price is posted, its pricing rule will be more complex. Such will be the case whenever there is a slow-moving element in the individual firm's expected demand, so

<sup>2</sup>Cases in which factor costs, and in particular labor costs, are affected by exchange rates through indexation have been extensively researched, especially in the context of open economies with staggered wage setting à la John Taylor (1980).

that its market share may be regarded as an investment. The customer market hypothesis, in which consumers slowly drift away when the price is raised, is one possible rationale for this intertemporal dimension; but there are other instances, like reputation. Then, not only current period's, but also future periods', expected profits depend on the present price decision, which becomes similar to any investment decision. The firm's maximization problem now involves discounting future profits; with perfect capital markets, this is done using the market interest rate of the relevant maturity. The individual supply schedules will thus be shifted up by an increase in the real interest rate, which can be regarded as the relative price of the firm's future real profits in terms of current ones (Fitoussi and Phelps). In order to capture this effect in the simplest possible way, the aggregate price equation may be rewritten as

$$(5) \quad q_t = \alpha' \cdot q_{t-1} + \varepsilon' \cdot E_t z_t + \eta' \cdot E_t x_t + \gamma \cdot R_t$$

where  $R_t$  is the expected real rate of interest that corresponds to the relevant time structure of the model, and is assumed to be a synthetic indicator of anticipated market conditions beyond the period for which individual prices are fixed.

### III. Macroeconomic Consequences

To briefly explore some of the macroeconomic implications of price setting in the open economy, the model has to be closed by specifying the process-generating aggregate demand in each country and exchange rate determination. In the case of two countries with floating exchange rate, ignoring the complications arising from consumers' and investors' expectations, the simplest, standard assumptions include perfect capital mobility and asset substitutability (i.e., interest rate parity), which may be written as

$$(6) \quad i_t = i_t^* + E_t x_{t+1}$$

where  $i_t$  and  $i_t^*$  are the home and foreign nominal interest rates, respectively, and  $x_{t+1}$

is the percent depreciation of the home currency between period  $t$  and period  $t+1$ , based on the average exchange rate prevailing during the period. Aggregate demand in each country is assumed to be influenced by domestic macroeconomic policies in the following way:

$$(7) \quad z_t = (m_t - q_t) + d_t$$

which may be considered as a simple reduced-form, IS-LM type of demand determination, where  $m_t$  is the percentage change in nominal money supply and  $d_t$  is a shift parameter representing fiscal policy. A similar equation is assumed to hold in the foreign country. The aggregate price index in each country is given by equation (5), which implicitly determines an aggregate supply curve.

As an illustration, let us consider the short-run responses of both economies to an unanticipated, sustained fiscal shock in the home country. To the extent that it raises the interest rate (with constant money supply), it will boost velocity in both countries, thus stimulating demand in both countries, a standard outcome in models featuring this kind of demand and exchange rate determination. With the domestic interest rate rising more than the foreign one, this will result in an instantaneous appreciation of the home currency and an anticipated depreciation, also a standard result. Foreign firms will tend to increase prices and markups in the home market in response to the currency exchange movement and the rise in world nominal interest rates, provided the latter were sufficient to raise anticipated real rates; however, they will not usually adjust to the full amount of the appreciation and will enjoy increased market shares. Prices and markups of domestic firms in the home market will go down, but by much less than the amount of the currency appreciation, as foreign prices in domestic currency unit increase, and higher real interest rates tend to inflate markups. In the foreign country, on the other hand, the exchange rate and the real interest rate effects work in the same direction to inflate prices and markups of firms based in that country, while their market shares also tend to increase there. On

As a whole, domestic firms' markups and market shares will tend to decrease on both markets, while foreign firms will experience variations in the opposite direction; but, in general, prices and markups of a given firm are likely to evolve differently in each market.

In the home country, the exchange rate effect on markups implies a tendency for the aggregate price index to decrease, though the expected demand and real interest rate effects both act as countervailing forces. In the foreign country, on the other hand, the aggregate price index will sluggishly rise in response to exchange and interest rate changes. Due to the corresponding tendency for real balances to be eroded there, the outcome will be a demand contraction if macroeconomic policies are not accommodating in that country, and even more so if foreign authorities actively oppose "imported" inflation, a result which generalizes the conclusions of the Fitoussi-Phelps analysis to the present context of "pricing to market." However, whether production increases or decreases in the foreign country depends on the precise magnitude of the relevant elasticities.

Analyzing the medium-term dynamics and the adjustment path of the variables in this stylized model would require an adequate specification of the model's time structure, of policy regimes, and of expectations formation, which is well beyond the scope of this paper. However, without actually solving for the long-run equilibrium of the system, we may hint at some of its most salient characteristics. In the present context, the appropriate conditions for long-run equilibrium ought to be that, with constant policies, expected exchange rate equals observed exchange rate, and that the trade balance be in equilibrium, so that there is no sustained capital flow. However, with the kind of price-setting behavior investigated here, insofar as changes in market shares depend on pricing behavior, which in turn depends on initial market shares, trade balance equilibrium will, in most cases, entail an exchange rate level that corresponds neither to purchasing power parity, nor to relative cost parity, even when countries are similar in every respect, except for the initial shares of markets held by domestic and foreign firms. In general, long-run equilibrium following a

shock is bound to depend on initial conditions and on the time path of adjustments, an outcome that arises in many imperfectly competitive settings and is made even more likely in this international environment with geographically segmented markets.

#### IV. Concluding Remarks

Simple hypotheses concerning market imperfections and price setting by firms can thus have far-reaching consequences in the context of open-economy macroeconomics. The conditions in which they arise are intuitively appealing, especially that of pricing to market; some of the results of the simple models featuring these assumptions seem to mimic available evidence, in particular on nominal and real exchange rates and on markups. But, owing to the extreme crudeness of the specifications discussed here, it is too early to decide whether the predicted patterns fit empirical data.

If they do have some validity, their major conclusion—namely, price setting in a context where firms price to market—has important macroeconomic implications. In the case of open economies with flexible exchange rates, the foregoing analysis suggests that a combination of integrated world capital markets and geographically segmented goods markets will often produce a great deal of instability, as measured by the variance of the economic aggregates, in response to shocks originating in policy changes or elsewhere, and this even if wages are highly flexible. In the case of countries managing their exchange rate, it also suggests that the choice of a target may not be self-evident; indeed, if some kind of purchasing power parity or relative cost parity policy is pursued, it may lead to sustained trade imbalances, a case that may characterize the current situation of countries participating in the European monetary system.

#### REFERENCES

- Akerlof, George A. and Yellen, Janet L., "A Near-Rational Model of the Business Cycle, with Wage and Price Inertia," *Quarterly Journal of Economics*, Suppl. 1985, 100, 823–38.

- Blanchard, Olivier J.**, "Price Asynchronization and Price Level Inertia," in R. Dornbusch and M. Simonsen, eds., *Indexation, Contracting and Debt in an Inflationary World*, Cambridge: MIT Press, 1983.
- Dornbusch, Rudiger**, *Open Economy Macroeconomics*, New York: Basic Books, 1980.
- \_\_\_\_\_, "Exchange Rates and Prices," *American Economic Review*, March 1987, 77, 93-106.
- Fitoussi, Jean-Paul and Phelps, Edmund S.**, *The Slump in Europe*, Oxford: Basil Blackwell, 1988.
- Helpman, Elhanan and Krugman, Paul**, *Market Structure and Foreign Trade*, Cambridge: MIT Press, 1986.
- Krugman, Paul**, "Pricing to Market when the Exchange Rate Changes," NBER Working Paper No. 1926, May 1986.
- \_\_\_\_\_ and **Baldwin, Richard E.**, "The Persistence of the U.S. Trade Deficit," *Brookings Papers on Economic Activity*, 1:1987 1-43.
- Le Cacheux, Jacques and Lecointe, François**, "Changes réels et compétitivité," *Revue d'OFCE*, July 1987, 20, 149-87.
- Phelps, Edmund S. and Winter, Sydney G.**, "Optimal Price Policy under Atomistic Competition," in E. S. Phelps et al., *Microeconomic Foundations of Employment and Inflation Theory*, New York: W. W. Norton 1970.
- Stiglitz, Joseph E.**, "Competition and the Number of Firms in a Market: Are Duopolies More Competitive than Atomistic Markets?," *Journal of Political Economy* October 1987, 95, 1041-61.
- Taylor, John B.**, "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, February 1980, 88, 1-23.

# Exchange Rates, Wages, and the International Allocation of Capital

By SLOBODAN DJAJIĆ\*

Divergence between macroeconomic policies of the major industrial countries over the last decade has generated massive swings in exchange rates and a seemingly endless string of growing deficits on the trade account of the United States. In the early 1980's, implementation of the Mundellian policy mix in the United States—a combination of tight monetary and expansionary fiscal policies—has produced the desired result of drastically reducing inflation and bolstering the value of the dollar, while simultaneously providing a fiscal stimulus to counter some of the contractionary effects of tight money on aggregate demand. Propelled by the sharp appreciation of the dollar, part of the stimulus flowed overseas. This benefited America's trading partners, especially those attempting to balance their federal budgets. However, an important consequence of a fiscal expansion in one country at the time of the fiscal consolidation elsewhere was a deterioration on the trade account of the expanding economy.

A move by the Federal Reserve in the direction of a more accommodative monetary policy over the last three years, and the market's steady loss of confidence in Washington's ability to bring the budget and trade deficits under control, have resulted in an extraordinary depreciation of the dollar against other major currencies. Although this depreciation has taken the dollar's value to new postwar lows, growth of the U.S. trade deficit has continued uninterrupted.

Lack of improvement on the trade front in spite of the exchange-rate adjustment presents not only a puzzle for economists, but also an important political problem. It is a source of friction in international negotia-

tions on macroeconomic policy coordination as well as a source of fuel for special interest groups trying to generate protectionist sentiment in the U.S. Congress.

A number of arguments have been advanced in an attempt to explain the persistence of the trade deficit. The most prominent among them is that based on the observation that dollar prices of goods imported into the United States adjust slowly in response to exchange-rate changes and that trade flows adjust slowly in response to price changes. Other explanations rest on the notion that the dollar has not fallen enough. They point to a range of factors which may have lowered over the last decade the level at which the foreign-currency value of the dollar is consistent with balanced trade. These factors include lagging productivity growth in the United States and the possibility that the strong dollar of the mid-1980's may have done persistent damage to the ability of American exporters to market goods abroad (Paul Krugman and Richard Baldwin, 1987).

This paper develops a model of international macroeconomic adjustment with an emphasis on certain relationships that may be helpful in shedding light on the deficit-persistence problem as well as on the increasingly serious unemployment problem confronting European economies. Central to the analysis is the link between the exchange rate and profitability in economies with rigid nominal wages (Pentti J. K. Kouri, 1979). An attempt is made to build around this link a simple model of an integrated world economy where exchange rates affect investment decisions of firms and where these decisions, in turn, influence the evolution of trade flows, exchange rates, wages, and economic activity across countries.

The analytic framework is presented in Section I. Section II utilizes this framework to discuss the implications of some of the

\*The Graduate Institute of International Studies, 132 rue de Lausanne, 1211 Geneva 21, Switzerland.

major macroeconomic policy shifts in the United States during the 1980's, both from the perspective of the U.S. economy and from that of its trading partners. Finally, Section III concludes the paper with brief remarks on future prospects for the U.S. trade deficit.

### I. The Model

Trade-account implications of exchange-rate changes are typically analyzed within the framework of a two-commodity world. One commodity is exported by the home country and the other by the foreign economy. Given the prevailing price of each commodity within the exporting country, depreciation of domestic currency lowers the relative price of the domestic good. The resulting substitution of domestic for foreign goods within the consumption baskets of both countries stimulates production activity at home while depressing it abroad. On the assumption that the associated changes in national incomes dominate the possible changes in national absorption, the trade account of the home country improves.

This familiar expenditure-switching effect of currency depreciation is fundamental to an understanding of the international adjustment mechanism. However, in an attempt to shift attention away from the usual explanations of the deficit-persistence problem and other issues related to the degree of substitutability between domestic and foreign goods, I shall proceed under the extreme assumption that the world economy produces a single commodity. While this strategy places limitations on the applicability of the model, it will prove very useful in focusing the analysis on certain exchange-rate effects which operate directly on the supply side.

Arbitrage in the market for the single commodity insures that the domestic price level,  $p$ , is equal to the foreign price level,  $p^*$ , adjusted by the exchange rate,  $e$ :

$$(1) \quad p = e + p^*.$$

All variables are in natural logarithms, except for the interest rates (defined below). Variables with an asterisk pertain to the

foreign country. Time subscripts are omitted for notational simplicity.

In both economies the demand for money is a function of output and the money rate of interest. Under the simplifying (and by no means essential) assumption of fixed-coefficients production technology, appropriate choice of units enables us to set output and employment of each country equal to the number of units of capital,  $K$ , located within its borders. Accordingly, the money-market equilibrium conditions for the two countries may be written as

$$(2) \quad m - p = \phi K - \beta i,$$

$$(3) \quad m^* - p^* = \phi^* K^* - \beta^* i^*,$$

where  $m$  is the nominal money stock,  $i$  is the nominal rate of interest, and Greek symbols are constant elasticities (and in some cases, below, speeds of adjustment) defined to be positive.

Using a dot over a variable to signify differentiation with respect to time, interest arbitrage, along with the perfect-foresight, risk-neutrality, and perfect-substitutability assumptions, implies that

$$(4) \quad \dot{e} = i - i^*.$$

Assuming that money demand functions are identical in the two economies (i.e.,  $\phi = \phi^*$  and  $\beta = \beta^*$ ), and recalling equations (1)–(3), we may express  $\dot{e}$  as

$$(5) \quad \dot{e} = (m^* - m + \phi k + e)/\beta,$$

where  $k \equiv K - K^*$ .

Turning to the markets for labor, let  $N$  and  $N^*$  represent the (logs of) labor supplies at home and abroad. Normally these supplies are increasing functions of the local after-tax wage rates. Assuming that the elasticity of labor supply,  $\alpha$ , is identical across countries, we may write

$$\begin{aligned} n &\equiv N - N^* \\ &= \tilde{n} + \alpha[(w - p) - (w^* - p^*) - t], \end{aligned}$$

where  $\tilde{n}$  is a constant. The expression in



brackets is the log of the ratio of domestic to foreign after-tax real wage rates;  $w$  is the log of the nominal wage and  $t$  is a policy variable which reflects the excess of the domestic over the foreign rate of labor-income taxation. Noting that  $p = e + p^*$  and defining  $v$  as  $w - w^*$ , we have

$$(6) \quad n = \tilde{n} + \alpha(v - e - t).$$

The demand for labor in each country is assumed equal to the number of units of capital located within the economy. In the event that the demand for labor exceeds the local supply, let us suppose that firms can satisfy their demand by compelling workers to provide additional (overtime) labor at the prevailing nominal wage.

While it is assumed that both  $w$  and  $w^*$  are predetermined at each instant, they may rise or fall over time at a rate proportional to a measure of disequilibrium in the corresponding labor market. Under the (admittedly unrealistic) assumption that the wage-adjustment process is symmetric in the two economies, we may write  $\dot{w} - \dot{w}^*$  as

$$(7) \quad \dot{v} = \mu(k - n) \\ = \mu[k - \tilde{n} - \alpha(v - e - t)].$$

Given the short- to intermediate-run perspective of the present model, the total capital stock of the world economy is taken as given. This stock is controlled by multinational firms and its international allocation is predetermined at each point in time. However, firms may gradually move capital from one economy to another if production costs differ across countries. Within the present model, the international production-cost differential is determined by the relationship between domestic and foreign real wage rates.

In addition to cost factors, the global allocation of capital is becoming increasingly responsive to growth of protectionist sentiment in the United States. Japanese firms, for example, are currently moving their operations across the Pacific, not only because real wages are relatively lower in the United States, but also due to fear of protectionism which is leading them to accept a new phil-

osophy: production should be where the market is (Lawrence Fisher, 1987). Moreover, for a growing number of leading Japanese and West German firms, the objective is not only to produce in the United States for the local market, but also to export from the United States to other markets. An attempt is made to capture these phenomena as simply as possible by assuming that

$$(8) \quad \dot{k} = \tau(e - v) + \sigma s,$$

where  $s$  is a variable (exogenous to the model) reflecting the level of protectionist sentiment in the home country, holding protectionism constant abroad. Alternatively,  $s$  may be interpreted as a fiscal policy parameter reflecting the rate at which investment is subsidized at home over and above the rate at which it is subsidized abroad.

Finally, goods-market equilibrium requires that the (constant) global output be equal to the sum of private and public expenditures throughout the world economy. Out of that relationship emerges the equilibrium real rate of interest.

## II. Solution and Implications of the Model

Equations (5), (7), and (8) constitute a system of three differential equations in  $k$ ,  $v$ , and  $e$ . The system can be shown to exhibit saddlepoint stability (see my 1987 paper). The solution for the exchange rate along the saddlepath is given by

$$(9) \quad e = \bar{e} + u_1(k - \bar{k}) + u_2(v - \bar{v}),$$

where  $u_1$  and  $u_2$  are constants such that  $u_1 < 0$  and  $1 > u_2 > 0$ . Variables with overbars are steady-state values. They are related to the exogenous variables as follows:

$$(10) \quad d\bar{e} = dm - dm^* - (\phi\alpha\sigma/\tau) ds + \phi\alpha dt,$$

$$(11) \quad d\bar{k} = (\sigma\alpha/\tau) ds - \alpha dt,$$

$$(12) \quad d\bar{v} = dm - dm^* \\ + [\sigma(1 - \phi\alpha)/\tau] ds + \alpha\phi dt.$$

Thus, in the context of the present model, a cut in the tax on labor income at home (i.e.,  $dt < 0$ ) causes domestic currency to appreciate in (what is from the model's perspective) the long run. In addition, it lowers domestic relative to foreign nominal wages and attracts capital from abroad. An investment subsidy or an increase in protectionism at home (i.e.,  $ds > 0$ ) has qualitatively the same long-run effects on the exchange rate and the international allocation of capital. As employment and the capital stock are directly related, the implications of tax cuts, investment subsidies, and protectionism for the levels of employment in the two economies follow immediately.

These findings are consistent with the generally accepted view that the U.S. tax acts of 1981 and 1982 have contributed to the appreciation of the dollar during the first half of this decade, as well as to the vigor of the U.S. economic recovery from the 1982 recession. That recovery has produced significant advances in employment in the United States, while job opportunities have become increasingly scarce in other industrial countries, particularly those belonging to the European Community. Although many observers attribute the stagnation in Europe in the 1980's to insufficient aggregate demand, the present model suggests that Europe's failure to follow the United States in offering incentives to stimulate investment and the supply of work effort may also be responsible. (For a much more comprehensive discussion of these and other significant factors contributing to the European slump, see Jean-Paul Fitoussi and Edmund Phelps, 1986, 1988.) The model also suggests that recent growth of protectionism in the United States—through its impact on the international allocation of capital—is likely to confront other major economies which rely heavily on exports, particularly Japan and West Germany, with an even slower pace of employment growth in the years to come.

A move by the Federal Reserve in the direction of a more accommodative monetary policy during the second half of this decade, in conjunction with restrictive monetary policies followed by other major industrial countries, has very similar short-term implications. As may be ascertained with the

aid of equations (9)–(12), an unanticipated increase in the domestic money stock (i.e.,  $dm > 0$ ) entails an instantaneous depreciation of domestic currency. The depreciation lowers domestic relative to foreign real wages, triggers an initial inflow of physical capital from abroad, and raises domestic at the expense of foreign employment. However, the "beggar-thy-neighbor" character of an expansionary monetary policy manifests itself only in the short run. In the long run,  $w$ ,  $e$ , and  $p$  all rise by as much as  $m$  does, while  $w^*$ ,  $p^*$ , and  $k$  return to their original levels. Accordingly, resistance of the West German Bundesbank in the fall of 1987 to calls for a looser monetary policy may be interpreted to reflect the bank's concern over the long-run implications of monetary expansion for the levels of wages and prices, and willingness to accept the transitory, adverse effect of a strong currency on investment and employment opportunities.

What are the trade-account implications of the measures considered above? A complete analysis of this problem requires a careful evaluation of how the measures affect the economy's income in relation to absorption. For our purposes it is sufficient to note that a tax cut on labor income, an investment subsidy, a rise in protectionist sentiment, or a monetary expansion, all contribute to a *transitory* trade deficit through at least one channel. They induce firms to temporarily increase investment spending at home and reduce it abroad in an attempt to reposition capital in favor of the home country. To the extent that some of these measures (tax cut and monetary expansion) stimulate private consumption at home in addition to investment, the initial trade-account deterioration is more pronounced. Thus, the puzzle of growing U.S. trade deficits can be resolved by noting that the 1985–87 fall of the dollar was primarily in response to current (and expected future) monetary expansion on the part of the Federal Reserve. Because this expansion also served to stimulate consumption and investment, it contributed to a transitory deterioration of the trade account.

These observations make me somewhat concerned over the recent tendency of currency-market participants to attach exces-

sive significance to the monthly trade figures in formulating exchange-rate expectations, while seemingly neglecting the evolution of the fundamental determinants of the exchange rate (the current and expected future paths of domestic and foreign money supplies and demands). If depreciation of the dollar and growth of protectionist sentiment does stimulate investment in the United States, causing a transitory trade-balance deterioration, and the latter is interpreted by the market as a signal that the dollar must fall further, a vicious deficit-depreciation cycle may develop. In the event that it does develop, the world economy will pay, once again, in terms of missallocated resources, for the excessive exchange rate movements (Rudiger Dornbusch, 1986, and Maurice Obstfeld, 1985). A firm commitment on the part of the major central banks to support the dollar seems warranted at this time.

### III. What are the Future Prospects for the U.S. Trade Deficit?

For several reasons, I expect the 1988-89 period to bring a very significant trade balance improvement. First, the reallocation of capital in favor of the U.S. economy implies future gains in U.S. employment and national income and relative stagnation abroad. Second, foreign governments are likely to respond at some point to their growing unemployment problem by implementing expansionary policies and measures to remove some of the existing structural limitations on growth. These policies would have a positive effect on foreign spending, as would the eventual return of Latin American debtor countries to the world market for investment goods. At the same time, expenditure growth

in the United States is likely to decline due to reductions in public spending and slower growth of private spending. As the ratio of U.S. to foreign national income rises and that of U.S. to foreign spending declines, the balance of trade will quickly improve.

### REFERENCES

- Djajić, Slobodan, "Exchange Rates, Wages, and the Allocation of Capital in the World Economy," mimeo., Graduate Institute of International Studies, Geneva, December 1987.
- Dornbusch, Rudiger, "Flexible Exchange Rates and Excess Capital Mobility," *Brookings Papers on Economic Activity*, 1:1986, 209-26.
- Fisher, Lawrence M., "Dollar's Fall, Protectionism's Rise Make U.S. Plant Sites Attractive," *International Herald Tribune*, November 28-29, 1987, 7.
- Fitoussi, Jean-Paul and Edmund S. Phelps, "Causes of the 1980s Slump in Europe," *Brookings Papers on Economic Activity*, 2:1986, 487-513.
- and ———, *The Slump in Europe*, Oxford: Basil Blackwell, 1988.
- Kouri, Pentti J. K., "Profitability and Growth in a Small Open Economy," in Assar Lindbeck, ed., *Inflation and Employment in Open Economies*, Amsterdam: North-Holland, 1979.
- Krugman, Paul R. and Baldwin, Richard E., "The Persistence of the U.S. Trade Deficit," *Brookings Papers on Economic Activity*, 1:1987, 1-43.
- Obstfeld, Maurice, "Floating Exchange Rates: Experience and Prospects," *Brookings Papers on Economic Activity*, 2:1985, 369-450.

# A Working Model of Slump and Recovery from Disturbances to Capital-Goods Demand in an Open Nonmonetary Economy

By EDMUND S. PHELPS\*

In my 1977 paper on wage indexation, it was commented that an implication of tying nominal wages fully to the consumer price index was that a shock acting to depress only the nominal prices of capital goods, not consumer prices, would fail to launch the adjustment of the wage level required to maintain "full" or (more generally) "equilibrium" employment of the inelastically supplied labor force. A velocity shock would leave money wages too high, a real shock would leave the real wage too high. Indexing in laws or agreements generally lets the wage level adjust only gradually, with the arrival of new workers.

In the 1988 monograph with Jean-Paul Fitoussi, this seed of an idea was planted in a complete model. In the third of our two-country models, stimulatory fiscal shocks in "America" raise the world real rate of interest and hence depress nontradable capital-goods prices in a stylized "Europe" where wages are fully indexed to the cost of living and slide down only in response to increased unemployment; the resulting fall in the amount of money demanded serves only to drive up the nominal consumer price index and the exchange rate—thus to raise nominal wages—not to avert the consequent slump of capital-goods employment and aggregate employment.

The essence of this argument, it became clear, is that real wages (defined here always in terms of consumer goods) are rigid, or at least sticky, while capital goods' real prices (likewise defined) are variable and jumpy. Nothing of importance is gained, if indeed

wages are highly indexed, and some transparency in the analysis is lost, by inserting money into the model. With the present paper I first refashion the Fitoussi-Phelps thinking along nonmonetary lines, and analyze more formally the implied dynamics. To facilitate the analysis I confine this model to that of a small open economy, small enough that it has only a negligible effect on the real interest rate determined in the world capital market. Two disturbances to the real demand price of capital are studied: a world real-interest shock and a marginal-efficiency-of-capital shock.

The radical nature of this exercise will not escape the reader. The available monetary macro models are passed up in favor of an entirely "real" theory of fluctuations in business output and employment—a New Keynesian real theory, it is important to note, in which positive real-interest and real-wage shocks are contractionary for the economy, not expansionary as typically portrayed in the neoclassical pages of non-monetary theory. It is, however, only a *working* model since the labor-market behavior it postulates is not wholly grounded on micro-theoretic foundations. Some micro-based formulations are underway.

## I. General Features

Output,  $Z_c$ , of the consumer good is subject to a constant-returns-to-scale production function,  $\phi$ , of the usual type:

$$(1) \quad Z_c = \phi(K, N_c).$$

Inada's derivative conditions,  $\phi_N(K, 0) = \infty$  and  $\phi_N(K, N) > 0$ , are used. Here  $N_c$  is the number of workers producing the consumer good.  $K$  is the totality of the capital stock as we suppose for maximum simplicity that

\*Department of Economics, Columbia University, New York, NY 10027. This paper was completed at the Seconda Università, Tor Vergata, Rome.

nly labor is used to produce the investment good. Letting  $N_I$  denote the number of workers producing the capital good and  $Z_I$  their output, we postulate

$$2) \quad Z_I = \gamma N_I, \quad \gamma > 0.$$

The market for each of these products is supposed always to clear, so the outputs are given by supply. The supply of these goods is a function of the real price,  $p$ , of the capital good, the real wage,  $v$ , and input constraints. As all firms can produce both goods, and all firms are alike, employees are everywhere allocated in such a way as to satisfy the first-order conditions for a maximum of consolidated gross profit,  $\phi(K, N - I_I) + p\gamma N_I - vN$ , where  $N$  is the total number of employees actively engaged, subject to the constraint that  $N$  cannot exceed the number available,  $L(v, r)$ , where  $r$  is the real interest rate, and to the nonnegativity constraint on  $N_I$ :

$$3) \quad \begin{array}{ll} \text{If } N_I > 0 & \text{If } N_I = 0 \\ \text{If } N < L & \text{inadmissible } \phi_N = v \geq p\gamma \\ \text{If } N = L & \phi_N = p\gamma \geq v \quad \phi_N \geq \max(v, p\gamma) \end{array}$$

Hence, the opportunity cost of  $N_c$  is  $v$  or  $p\gamma$ , whichever is greater, and  $\phi_N$  is driven down to that cost unless both constraints are binding, leaving  $N_c$  with no slack for increase.

For simplicity, the analysis will focus on the case in which  $L(v, r)$  is a constant,  $L$ . We shall impose the welcome restriction that the initial real wage is high enough in relation to the capital stock that the  $N_c$  "demand" level, say  $\tilde{N}_c$ , equating  $\phi_N(\cdot)$  to  $v$  is less than  $L(v, r)$ . Without that restriction a shock that drives  $N_I$  to zero will still leave a positive excess total demand for labor. At least at first, then, the lower right-hand possibility does not apply, so that  $\phi_N = \max(v, p\gamma)$  and  $N_c$  is given by the inverse  $\phi_N^{-1}(K, \max(v, p\gamma))$ . In fact, as will be clear, if the economy is initially in its steady state, then  $p\gamma = v$  initially, with  $N_I > 0$  and  $N = L$ .

For all  $(K, L, v)$  satisfying the restriction, the two "supply curves," measured in units of full-time labor rather than output, are those shown in Figure 1. The total number of active jobs supplied is given by the dashed

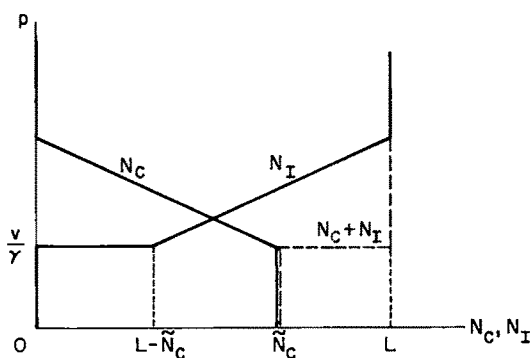


FIGURE 1. SUPPLY CURVES IN TERMS OF LABOR INPUT

curve. In equational form, the supply of the capital good is given by

$$(4) \quad Z_I = \begin{cases} 0 & \text{if } p\gamma < v. \\ \gamma [L(v, r) - \phi_N^{-1}(K, v)] & \text{if } p\gamma = v. \\ \gamma [L(v, r) - \phi_N^{-1}(K, p\gamma)] & \text{if } p\gamma > v; \end{cases}$$

and total active employment is given by

$$(5) \quad N_I + N_c = \begin{cases} \phi_N^{-1}(K, v) & \text{if } p\gamma < v. \\ L(v, r) & \text{if } p\gamma = v. \\ L(v, r) & \text{if } p\gamma > v. \end{cases}$$

We shall see that when the above restriction holds initially it may continue to hold along the postshock path found below in which case we need not replace (4) and (5) with their (more cumbersome) unrestricted versions.

The following dynamic equations can now be understood, taking as given for the moment the path of the real interest rate. The dynamics of the capital stock come from exponential depreciation at rate  $\delta > 0$  and the supply of  $Z_I$  in (4).

$$(6) \quad \dot{K} = \begin{cases} -\delta K & \text{if } p\gamma < v. \\ -\delta K + \gamma [L(v, r) - \phi_N^{-1}(K, p\gamma)] & \text{if } p\gamma \geq v. \end{cases}$$

We sometimes write  $\phi_N^{-1}(\cdot)$  as  $l(\phi_N)K$  where  $l(\cdot)$  is the  $N_c/K$  ratio,  $l'(\cdot) < 0$ .

The dynamics of the real price of capital comes from the arbitrage condition,

$$(7) \quad \dot{p} = (r + \delta)p - R^c(\max(v, p\gamma)), \\ R^{c'}(\phi_N) < 0,$$

where  $R^c$ , the real rental on capital, can be seen to equal  $\phi_K$  when  $\phi_N = \max(\nu, p\gamma)$ , and therefore bears the familiar factor-price-frontier relation to  $\phi_N$  and hence to  $\max(\nu, p\gamma)$ . In identifying the actual price change with the expected change, I am taking the economy to be on an equilibrium trajectory.

For present purposes I shall assume that the dynamics of the average real wage is described by Samuelson's gradualist version of supply and demand. If  $N_I$  is less than the partial-excess supply  $L(\nu, r) - \phi_N^{-1}(K, \nu)$ , the average wage will be falling. If  $N_I$  is greater, so that in fact  $\phi_N = p\gamma > \nu$ , the wage will be rising. And if  $N_I$  is equal, the wage will be unchanging. Using (2) and (4), and a constant speed of adjustment coefficient  $\alpha > 0$ ,

$$(8) \quad \dot{\nu} = \alpha \begin{cases} 0 - [L(\nu, r) - \phi_N^{-1}(K, \nu)] & \text{if } p\gamma < \nu \\ L(\nu, r) - \phi_N^{-1}(K, \nu) & \text{if } p\gamma = \nu \\ -[L(\nu, r) - \phi_N^{-1}(K, \nu)] & \text{if } p\gamma > \nu \end{cases}$$

Simplifying, we have

$$(8') \quad \dot{\nu} = \infty \left[ \phi_N^{-1}(K, \nu) - \begin{cases} L(\nu, r) & \text{if } p\gamma < \nu \\ \phi_N^{-1}(K, p\gamma) & \text{if } p\gamma \geq \nu \end{cases} \right]$$

The first term is  $N_c$  "demand" and the second term is  $N_c$  "supply" after netting  $N_I$  from  $L(\nu, r)$ . We suppose that the excess demand  $\phi_N^{-1}(K, \nu) - L(\nu, r)$  is decreasing in  $\nu$ , like  $\phi_N^{-1}(K, \nu)$ .

In a stationary state, then, if such exists,

$$(8) \quad \bar{\nu} = \bar{p}\gamma$$

$$(7) \quad \bar{p} = R^c(\bar{\nu})/(\bar{r} + \delta)$$

$$(6) \quad \bar{K} = [\delta + \gamma l(\bar{\nu})]^{-1} \gamma L(\bar{\nu}, \bar{r}),$$

$$l(\bar{\nu}) \equiv \bar{K}^{-1} \phi_N^{-1}(\bar{K}, \bar{\nu}),$$

where, again,  $l(\phi_N)$  is the  $N_c/K$  ratio,  $l'(\cdot) < 0$ .

## II. A Small Open Economy

Let us take the consumer good of my open economy to be tradable in a perfect world good(s) market. As the capital accumulation equation can be seen to imply, the domestically produced investment good is nontradable. The model could be interpreted as having in the background some imported capital goods as well that can be instantaneously hired at a world market (real) rental that remains fixed for all time.

To close the model let us suppose that the expected real rate of interest is given by the *ex ante* world real-interest rate, and the latter is a constant,  $r^*$ , over the indefinite future.

$$(9) \quad r = r^* \equiv \text{constant} > 0.$$

With  $r$  thus a parameter, equations (6)–(8') form a complete dynamic system.

### A. A Real Interest Shock

I study here the adjustment to a world interest shock in the form of an abrupt upward shift of the parameter  $r^*$  when the economy is initially at rest. Figure 2 is the key phase diagram. There the ray, which depicts the equation  $\nu = p\gamma$ , is the locus of points at which the excess supply of labor is zero, hence  $\dot{\nu} = 0$ ; to the right and below,  $N_I = 0$  so, unless our appealing restriction that  $\phi_N^{-1}(K, \nu) < L(\nu, r)$  should cease to hold, there is excess supply, hence  $\dot{\nu} < 0$ ; to the left and above, we would have excess demand and therefore  $\dot{\nu} > 0$ . The curve graphs  $p = R^c(\max(\nu, p\gamma))(r^* + \delta)^{-1}$ , and gives  $p$  as a decreasing function of  $\nu$  where  $\nu > p\gamma$  and is flat elsewhere. This curve is the locus of points at which  $\dot{p} = 0$ ; above the curve,  $\dot{p} > 0$ , and below,  $\dot{p} < 0$ . It follows that  $p$  and  $\nu$  have a uniquely determined saddlepath solution given by the broken curve. The intersection of these curves corresponding to the new  $r^*$  gives the new  $\bar{p}$  and  $\bar{\nu}$ .

The initial preshock state was at  $(p_0, \nu_0)$  lying on the ray somewhere above the new  $(\bar{p}, \bar{\nu})$  since an increase of  $r^*$  lowers the

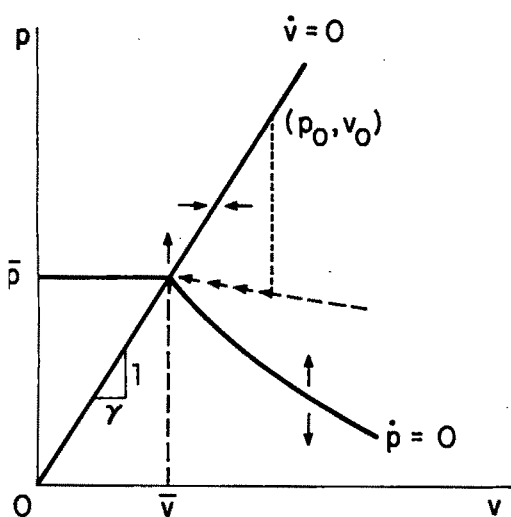


FIGURE 2. PHASE DIAGRAM FOR THE SMALL OPEN ECONOMY

$\dot{v} = 0$  locus and so decreases  $\bar{p}$  and  $\bar{v}$ . At the moment of the interest shock, the saddlepath, which must have passed through  $(p_0, v_0)$  prior to the shock, is also lowered to the position shown by the dashed line. The unique solution for the trajectory is a drop of  $p$  from  $p_0$  onto the saddlepath, whereupon  $p$  recovers monotonically to the reduced  $\bar{p}$  and  $v$  falls steadily to  $\bar{v}$ , reaching this point in finite time and thus marking the end of phase 1. Throughout this phase,  $K$  is steadily falling as the drop of  $p$  below  $\gamma^{-1}v$  shuts down the capital-goods industry. The decline of the capital stock, far from threatening to revoke the restriction  $L(v, r) > \phi_N^{-1}(K, v)$  which held initially, only aggravates the slack labor that  $N_c$  leaves, namely  $L(v, r) - \phi_N^{-1}(K, v)$ . This slack is still positive, I shall assume, and hence  $v$  is still too high to equate  $\phi_N^{-1}(K, v)$  single-handedly to  $L(v, r)$  even when  $v$  hits  $p\gamma$ ; so the restriction is never violated in this case.

At the start of the second phase,  $v$  and  $p\gamma$  having simultaneously reached the level  $\bar{p}\gamma$ , the capital-good industry springs back into operation.  $N_f$  jumps from zero to take up the whole slack left by the consumer-good industry,  $L(v, r) - \phi_N^{-1}(K, v)$ , and thus to

extinguish the excess supply prevailing in phase 1. Hence  $v$  is no longer falling. Nor does  $v$  begin to rise: Since  $v$  does not fall further,  $R^c(v)(r + \delta)^{-1}$  does not rise more, remaining at  $\bar{p}$ , so the only nonexplosive course for  $p$  satisfying (7) is to remain at  $\bar{p}$  in view of the saddlepoint instability of  $p$  around the saddlepath implied by  $\partial[p - R^c(p\gamma)(r + \delta)^{-1}]/\partial p > 0$  and indicated by the diverging vertical arrows around the  $\dot{p} = 0$  locus; since  $p$  does not rise above  $\bar{p}$ , causing  $p\gamma > v$ , no excess demand arises to pull  $v$  above  $\bar{v}$ . Thus  $(\bar{p}, \bar{v})$  is an absorbing state.

With  $p\gamma = v = \bar{v}$ , equation (6) implies that

$$(10) \quad \dot{K} = \gamma L(\bar{v}, r) - [\gamma l(\bar{v}) + \delta] K.$$

Hence  $K$  converges to  $\bar{K}$ . Note that since  $l(v)K - L(v, r)$  is taken to be decreasing in  $v$ , at least for all relevant  $K$  and  $r$ , the decrease of  $\bar{v}$  resulting from the interest shock tends to generate a decline in  $\bar{K}$ . (If  $L_r > 0$ , which I have ignored until now, that effect tends to work in the other direction.) Hence, though  $K$  arrives preshrunk by phase 2, it may shrink more.

#### B. A "Marginal Efficiency of Capital" Shock

Let us now study the adjustment of this nonmonetary economy to a marginal efficiency shock caused by a disturbance to expected future capital rentals: There is an anticipated drop in the size of some jump in the rental-wage relation,  $R^c(v)$ , that has for some time been in prospect at a certain future date, due to some impending technical innovation perhaps; and expectations of future rentals, being correct, are revised accordingly. I illustrate with the limiting case in which the *whole* of the future boost to rentals is suddenly "written off."

When the prospect of the future shift of the rental relation first arose,  $p$  must have jumped from its saddlepath level to a higher level in anticipation of the still higher  $p$  that would be sustained by the eventual shift (and the corollary shift of the saddlepath).

Thereupon, with  $p$  thus elevated,  $p$  must have been steadily rising to provide a rate of return (including capital gain) equal to the unchanged  $r^*$ , thus to satisfy (7). More important, if we suppose that the economy had been at the rest point,  $(\bar{p}, \bar{\nu})$ , when the prospect of the shift arose, the rise of  $p$  must have produced  $p\gamma > \nu$ , hence excess demand for labor, so that  $\nu$  must have rising above its previous rest-point level,  $\bar{\nu}$ . In a diagram like Figure 2, the path of  $(p, \nu)$  must have lain *above* the  $p\gamma = \nu$  ray, heading "north-east" toward its scheduled rendezvous with the new saddlepath path that will apply when the prospective shift of the  $R^c(\nu)$  relation arrives.

The sudden end to expectations of an improved rental-wage relation eliminates the prospective further rise of  $p$  on which its recent elevation above the saddlepath depended. Hence  $p$  immediately drops to its saddlepath level. But  $\nu$  by this time exceeds  $\bar{\nu}$ , owing to the aforementioned excess demand for labor. Hence the point on the saddlepath to which the equilibrium  $(p, \nu)$  drops is to the right of the saddlepath terminus, at  $(\bar{p}, \bar{\nu})$ , and hence *below* the  $p\gamma = \nu$  ray; therefore  $p\gamma < \nu$ , the capital-goods branches at all the firms suspend operations, and aggregate employment falls sharply. As earlier, recovery from the slump develops as the average real wage rate is reduced in response to the excess supply of labor.

If, instead, the economy had been at a saddlepath points to the right of the rest point,  $(\bar{p}, \bar{\nu})$ , when the prospects of the shift arose, hence in a slump to begin with, the prospective shift must have revived  $N$  to the full-employment level  $L$  if large enough in relation to the initial excessiveness of the wage to drive  $p$  *above* the  $p\gamma = \nu$  ray. In that case, the drop of the equilibrium  $(p, \nu)$  to the saddlepath again causes  $p$  to fall *below*

the ray; as before, then, employment vanishes in the capital-good branches.

Thus I have obtained a seemingly Keynesian downturn, sparked by a weakening of investment demand, in a model that is not Keynesian in the usual sense, and not even monetary. In fact, open-economy Keynesian models assert that, with flexible exchange rates, exports would "crowd in" to replace investment spending and thus fend off the slump. So the model here is more "Keynesian" than is Keynes.

### III. Comments

This paper has investigated a non-Keynesian world in which saving and investment shocks disturb the real prices of capital goods, through the real rate of interest or the marginal efficiency of capital, with effects on aggregate employment. The small open economy dynamic model here demonstrates anew the proposition of Fitoussi-Phelps, that capital-goods output at home might be crowded out by fiscal stimulus abroad. The model, however, suggests a wealth of ideas for a real theory of fluctuations in business employment.

### REFERENCES

- Fitoussi, Jean-Paul and Phelps, Edmund S., *The Slump in Europe: Reconstructing Open-Economy Theory*, Oxford: Basil Blackwell, 1988.
- Phelps, Edmund S., "Index Issues," in K. Brunner and A. Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conference on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 149-67.
- Samuelson, Paul A., *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1947.



## SEARCH BEHAVIOR IN LABOR AND PRODUCT MARKETS<sup>†</sup>

### Pareto Inefficiency of Market Economies: Search and Efficiency Wage Models

By BRUCE GREENWALD AND JOSEPH E. STIGLITZ\*

Serious macroeconomists have long been faced with a dilemma: how can one reconcile the seeming inefficiencies associated with the periodic episodes of unemployment and under utilization of capital with those rational, competitive forces which, in our traditional microeconomic paradigm, we argue ruthlessly seek out profitable opportunities, eliminate waste, and weed out incompetent producers. In this quest, economists have identified a number of ways in which our economy differs from the idealization of the Arrow-Debreu model that can explain the existence and persistence of unemployment, among the most important of which are search costs and the dependence of productivity on wages (the efficiency wage hypothesis). Once we recognize the importance of these, then the existence of unemployment need not be evidence of market inefficiency: economic efficiency requires the movement of labor from one job to another, as disturbances change the marginal productivity of workers in different industries; search takes time and resources; even if it were always *feasible* to move labor instantaneously from a low to a high productivity use, with no interim period of unemployment, it may—for some individuals, under some circumstances—be inefficient to devote the resources to search required for such transitions; it may be more efficient to spend a period unemployed. Indeed, the very words we use to describe the resulting unemploy-

ment rate, “the natural rate,” suggests that there is nothing particular perverse, or inefficient, about this unemployment.

By the same token, if productivity is increased by increasing wages, it is quite plausible that efficiency entails wages at above market-clearing levels.

More broadly, the approach taken by modern macroeconomists, in which the terms of the contracts between workers and employers takes into account not only the absence of income insurance for workers, but also search/mobility costs and efficiency wage considerations, seems to preclude the possibility that any resulting unemployment is inefficient: for the contracts are designed to be “locally efficient,” that is, to maximize the firm’s expected profits, given the reservation utility levels of workers.

The line of reasoning that we have presented in the preceding paragraph, as persuasive as it may seem, is simply wrong. The fundamental question in which we are interested is, is a decentralized market economy, characterized by search costs, efficiency wages, incomplete insurance markets, or by a variety of other informational imperfections, or that deviates from the standard specification of the competitive model in other ways that seemingly enhance its realism—is such an economy Pareto efficient? In judging the efficiency of the resulting market allocations, we need to take explicitly into account the costs of search or information acquisition; of the factors that make productivity dependent on wages; of the absence of a complete set of insurance markets. We ask, are there feasible government interventions, that respect these aspects of actual market economies, that can make everyone better off. (We do not ask, is it reasonable to

<sup>†</sup>*Discussants:* Dale T. Mortensen, Northwestern University; Kenneth Burdett, Cornell University.

\*Bell Communications Research Labs, Morristown, NJ 07960, and Princeton University, Princeton, NJ 08544, respectively.

assume that governments will actually intervene in such a way as to effect a Pareto improvement?) In deference to common usage, when there exist such interventions, we say that the economy is *constrained Pareto inefficient*; in adopting this language, we emphasize that we do not believe that the considerations under examination here, such as information costs, are any less "real" than production costs.

We show here that (for rather different reasons) market economies with search and efficiency wages are, in general, not constrained Pareto efficient. In our earlier work (1986), we proved a general theorem establishing that markets with imperfect information and incomplete markets were constrained Pareto inefficient. An explicit assumption of that analysis, however, was that markets cleared, whereas here we are concerned with situations where markets may not. Though efficiency may indeed entail the presence of some unemployment and that wages are not set at market-clearing levels, there is a presumption that neither the level of unemployment or wages is Pareto efficient.

### I. Efficiency Wage Models

The basic hypothesis of the efficiency wage model is that workers' productivity depends on the wage paid; here we generalize the standard formulation by allowing productivity (per hour) to depend also on the number of hours worked. Assume that there are  $L$  identical workers. The  $i$ th firm's output is simply a function of its effective labor supply,  $L_i h_i \Gamma_i(v_i, h_i)$  where  $v_i$  is the wage its workers receive (which may differ from the wage the firm pays,  $w_i$ , because of taxes) and  $h_i$  is the number of hours each of its  $L_i$  workers works:

$$(1) \quad Q_i = F_i(L_i h_i \Gamma_i(v_i, h_i)), \quad F_i' > 0,$$

$$\Gamma_1 > 0, \text{ and } \Gamma_{11} \leq 0 \text{ as } w \geq \hat{w}$$

The firm maximizes its profits,  $\pi = p_i Q_i - w_i h_i L_i$ , subject to the constraint that it must offer a contract that exceeds workers' reservation utility:

$$(2) \quad U(w_i, h_i; q) \geq \bar{U}$$

where utility is a function of wages and hours, as well as the consumer price vector,  $q$ . It is by now well known that the solution may entail the constraint (2) not being binding. We focus on this regime here. The maximized level of profits will be a function of prices and the relationship between wages paid and wages received; with an ad valorem wage tax,  $v = w(1 - \tau)$ , and we write:  $\pi_i^* = \pi_i^*(p_i, \tau)$ , with the standard result that the derivative of profits with respect to price is equal to the firm's output.<sup>1</sup> (Because wages are set by the firm, they do not appear explicitly in the profit function.)

The fact that wages may exceed market-clearing levels in equilibrium implies that we will need to divide consumers into two groups, the employed and the unemployed. Given consumer prices,  $q$ , the level of income (in excess of wage income, if any) required by an individual to attain a level of utility  $U^*$  is given by the modified expenditure functions:  $E^{ju} = E^j(q, 0, 0, U^*)$  for an unemployed household, and  $E^{je} = E^j(q, h_j, v_j, U^*)$  for an employed household working  $h_j$  hours and receiving a wage of  $v_j$  per hour.

The  $j$ th household owns a fraction  $a_{ij}$  of the  $i$ th firm. If the government imposes a set of taxes that changes  $p$ ,  $q$ ,  $h$ , or  $v$ , then for the  $j$ th household to attain utility level  $U^*$  requires a compensatory payment of  $\Delta E^j - \sum a_{ij} \Delta \pi_i^*$ , where  $\Delta \pi_i^*$  is the change in profits. We denote these compensation by  $I^j$ .

Assume the government imposes a set of commodity taxes, so the  $k$ th consumer price is now  $q_k = p_k + t_k$ ; an ad valorem wage tax at the rate  $\tau$ , and a tax per employed worker at the rate  $\mu$ . The profit function can be modified in a straightforward way to reflect the per employee tax, to read  $\pi_i^* = \pi_i^*(p, \mu, \tau)$ . Now, if the government can impose a set of taxes that raises revenue, after paying all individuals compensation that allows them to remain at the same level of utility they had attained in the market equilibrium, then the market equilibrium

<sup>1</sup>With the caveat that if productivity depends on consumer prices, then there is an additional term, reflecting the effect of the change in producer prices on consumer prices, and the effect of that on productivity, at any given level of wages and hours.

cannot have been (constrained) Pareto efficient. Government revenue is  $R = \sum t_k Q_k + \tau \sum w^j h^j + \mu L - \sum I^j$ , where  $L$  is aggregate employment,  $Q_k$  is aggregate consumption of the  $k$ th commodity, and where prices are determined at the market-clearing levels (with firms choosing their profit-maximizing levels of inputs and outputs, and households choosing their utility maximizing consumption bundles, constrained, of course, by the availability of jobs). Wages are set at profit-maximizing levels. For simplicity, for the remainder of the paper, we assume  $h$  is fixed.

Straightforward differentiation, making use of the standard properties of expenditure and profit functions, establishes that at  $t_i = 0, \mu = 0, T = 0$ ,<sup>2</sup>

$$dR/dt = \{E^{ju} - E^{je}\} dL/dt - L \{dE^{je}/dw\} \{dw/dt\}.$$

Similar expressions hold for changes in  $\tau$  and  $\mu$ . We decompose the total effects of the tax into four elements:

(i) A direct effect in raising consumer prices and government revenue. These are simply transfer effects—when the government compensates the individual for the increased prices, the two effects (for small taxes) cancel.

(ii) A general equilibrium effect on prices; an increase in prices raises profits, and lowers consumers' utility; again this is a transfer effect, and so long as the goods' market clears these effects cancel (recalling that every firm must be owned by someone, i.e.,  $\sum a_{ij} = 1$ ). (If productivity depends on consumer prices, then there is an additional, nontransfer, effect, from any change in consumer prices, equal to  $\sum p F' L h \Gamma_q \cdot (dq/dt)$ .)

<sup>2</sup>This expression holds if all firms are identical and all individuals (*ex ante*) are as well. More generally we write, for small taxes

$$\Delta R \approx \sum [\delta^j \{E^{ju} - E^{je}\} - (dE^{je}/dw) \Delta w_j]$$

where  $\delta^j = 1$  for a worker who was unemployed before the imposition of the tax and is employed after;  $-1$  for a worker who was employed before the tax and is unemployed after; and 0 otherwise.

(iii) An indirect effect on the profit-maximizing level of employment; by the envelope theorem, the effect on profits is zero, but the effect on consumers—since there is job rationing—is positive; the dollar value of this is equal to the difference between the compensation, net of wages received, required for the unemployed to be at the same level of utility as the employed. Because private firms ignore this term, market equilibrium entails too little employment.

(iv) An indirect effect on the wage level. Again, by the envelope theorem, the effect on profits is zero, but the effect on consumers' is positive (if wages increase.) Thus, there is a presumption that market wages are too low, even though they are set at above market-clearing levels.

Notice that this formulation not only establishes that there are welfare-enhancing government interventions, but also tells us precisely what kinds of interventions are desirable: those that increase employment and wages. Thus, a small ad valorem wage subsidy, that, at least in the simplest versions of the efficiency wage model, will leave consumer wages unchanged, will increase employment and hence increase welfare. Assume productivity is positively affected by food consumption and negatively affected by alcohol consumption, in such a way that the firm responds to a food subsidy and an alcohol tax by increasing employment, but leaving wages unchanged or increased; in these circumstances a food subsidy and an alcohol tax may be desirable.

## II. Search

It has long been recognized that search can give rise to unemployment, particularly if (at least for some individuals) off-the-job search is more efficient than on-the-job search. Although some search unemployment will then clearly characterize market equilibrium, it is again by no means clear that the level of unemployment will be Pareto efficient. We show that it is not, using a framework similar to that employed in our discussion of efficiency wages. Again, there will be employed and unemployed workers, now depending upon which workers successfully obtain jobs. Firms' decisions concern-

ing hiring, layoffs, and search and workers' decisions concerning quits and search intensities all generate "search" externalities, affecting the likelihood of a firm finding a well-matched worker and a worker finding a well-matched job.

To see the parallel with the earlier section as clearly as possible, we focus on a special case where all individuals and firms are (*ex ante*) identical, and where, in equilibrium, all firms pay the same wage. The probability of a match is  $\Phi(x, y)$ , where  $x$  is the vector of workers' search intensities,  $y$  vector of firms' "hiring" intensities. Employment,  $L$ , is just equal to  $N\Phi$ , where  $N$  is the number of potential workers. For simplicity, we partition the vector  $x = \{x_i; x^*\}$  where  $x^*$  is the search intensity of all other workers. Firms choose wages and hiring intensities to maximize expected profits (taking into account the effect of those decisions on the likelihood of a match); and their maximized value of profits can be represented by  $\pi_i^*(p; \tau, \mu; z_i)$ , where  $z_i$  is a description of the relevant market environment, here, the wages and hiring intensities of all other firms and the search intensities of all individuals. As before, we can write the expenditure function of those who are successful in obtaining a job and those who are not by  $E^{je}$  and  $E^{ju}$ , respectively, noting now the dependence on the market environment,  $z_i$ , which now includes the search intensities of others as well as firms' hiring and wage policies.

An identical argument to that employed before shows that if the government can impose taxes which raises revenue, after compensating individuals, then the market equilibrium is not constrained Pareto efficient. Again, straightforward differentiation yields

$$\begin{aligned} dR/dt \approx & \{E^{ju} - E^{je}\} [dL/dt - N\Phi_x(dx/dt)] \\ & - L \{dE^{je}/dw\} \{dw/dt\} \\ & + \Sigma (\partial \pi_i^* / \partial z^*) (dz^*/dt). \end{aligned}$$

The first term is slightly modified from its earlier form, to reflect the fact that the individual, in deciding on her or his search intensity, takes into account the expected gain

in utility from the increased likelihood of employment from additional search; the individual does not take into account the effect of those search decisions on the employment prospects of others, and firms do not take into account the gain in utility of those who do obtain jobs as a result of their increased recruitment activities.

There are two additional terms besides those discussed in the previous section, arising from the "external" effects on profits. An increase in hiring intensity by one firm reduces the likelihood of a match and hence has a negative effect on profits of other firms; similarly for wage changes. (These are however, total general equilibrium derivatives, and the indirect effect of these perturbations on workers' search intensity, and of that on profits, needs to be taken into account.)

The market failure we have identified here can be given a "missing markets" interpretation. Suppose there is a notional employment agency that pays  $q_x$  for search intensity  $x$  and  $q_y$  for hiring intensity  $y$ , and in turn receives payments of  $q_0$  for matches. The expected number of matches  $N$  is a function of the vector  $\{x^j, y^j\}$ . Then the employment agency maximizes  $q_0 \Sigma \Phi_j(x, y) - q_x x - q_y y$ . The function  $\Sigma \Phi$  looks like the production function of the employment agency. Since this formulation eliminates the externality, the solution to this problem in conjunction with the maximization problem of households and firms yields the Pareto optimal set of outcomes. Looking at the resulting equilibrium prices paid to the notional employment agency, we obtain the optimal taxes and subsidies that a government would have to impose on search-related activities in the absence of such an agency. And the degree to which the pseudo-production function  $\Sigma \Phi$  exhibits decreasing, increasing, or constant returns to scale determines whether these payments will leave a net surplus or deficit.<sup>3</sup>

<sup>3</sup> It is clear that if, upon each transaction, any surplus is divided among the participants, there is no division rule which will result in a Pareto efficient outcome unless the pseudo-production function exhibits constant returns to scale.

It is easy to generalize the results of this model, for instance to implicit contract models, where firms sign contracts with workers to maximize their profits, for a given (reservation) expected utility of workers. The contract will specify firms' retention, hiring and wage decisions as a function of the state of nature  $\Theta$ ; state contingent profit and expenditure functions can again be presented as a function of the market vector  $z$ ; and a state contingent tax on some commodity  $i$  is desirable if

$$dR/dt_i = \{\Sigma \pi_z^* - \Sigma E_z\} \cdot \{dz/dt_i\} \neq 0,$$

where the subscript  $z$  denotes a derivative with respect to  $z$ . A tax that discourages an individual from searching (say, because it increases the opportunity cost of searching) has positive externalities on other individuals, since, at any fixed level of search intensities on their part, it increases the likelihood that they will find a good (better) job. A tax that encourages firms to search more (by subsidizing new employees) or to retain more workers has positive externalities on workers, since at any fixed level of search intensities on the part of worker, the likelihood that they find a (better) match is increased, but negative externalities on other firms (because of the reduced likelihood of a match.) We conjecture, but have not proved, that normally the first effect dominates the second: there is too little hiring.

Notice that firms, in setting their lay-off rates, take into account the effect of changes in the lay-off rate of the expected utility of its own workers, but not the external effect of the search efforts of its workers on the likelihood of others' obtaining employment.

Such a tax may have a second set of effects on the wages offered by different firms; if firms change their wages for new hires, in response to the changed search intensity, there is a second-order effect on

profits, which can be ignored, but a first-order effect on workers' expected utility.

### III. Concluding Remarks

In our earlier work, we showed that market equilibrium with competition, in contexts in which all markets clear, but in which there was imperfect information or incomplete markets would not, in general, be Pareto efficient.

Here we have extended those results to incorporate equilibria in which firms are wage setters rather than wage takers, where they set their wage to take into account efficiency wage considerations (including the effect on the cost of recruiting workers and on labor turnover), and where they may set the wage at a level where markets do not clear. We believe that this provides a more accurate characterization of labor markets than is provided by the standard perfect information, market-clearing model.

It should be clear that similar results obtain in other contexts—in labor, product, and capital markets—in which wages, prices, and interest rates affect market behavior, for instance by conveying information. Though efficiency may indeed entail unemployment, credit rationing, or prices exceeding marginal costs of production, there is no presumption that the extent of rationing, and the level of wages, prices, and interest rates in the market equilibrium are efficient. The precise nature of the distortions depends on the exact specification of the model: in the efficiency wage model, there was too little employment, as firms failed to take into account the potentially large discrepancy in utility of the employed and the unemployed.

### REFERENCE

- Greenwald, B. and Stiglitz, J. E., "Externalities in Economies with Imperfect Information and Incomplete Markets," *Quarterly Journal of Economics*, May 1986, 101, 229–64.

# "High-Low Search" in Product and Labor Markets

By STEVE ALPERN AND DENNIS J. SNOWER\*

This paper is about commonplace activities of the following sort: (A) A firm makes a production decision in the face of an uncertain product demand. It gains demand information by putting a particular quantity of goods up for sale at a specified price and observing how much of this quantity is sold. This information, in turn, is used when it makes its next production decision. (B) A worker sets his or her wage in the absence of full information about his productivity at his firm. He gains productivity information by observing whether the firm offers him a job at the wage he has set. He uses this information in formulating his subsequent wage claims.

Although such activities in the product and labor markets are frequently encountered in practice, they are not captured by traditional economic theories of search, because of two important features:

(i) Agents engage in these activities not simply to maximize their welfare on the basis of their current information, but also to improve their information sets. The latter involves learning through participation in the market; in particular, it is about learning from the market outcomes of price or quantity decisions. We call this "learning from experience."

(ii) The information that agents gain through market participation characteristically contains what we call "high-low" elements. By observing the transactions that follow from their price-quantity decisions, agents generally do not gain enough information to determine whether these decisions were "correct" (i.e., whether these decisions make the agents as well off as possible, given

the actual market conditions). Rather, agents may merely discover whether their prices and quantities were "too high" or "too low". For instance, when the worker in example B above receives no job offer, he infers that he must have set his wage too high; if he is hired, he infers that his wage must have been too low or correct. The process of gathering high-low information we call "high-low search."

Traditional economic theory ignores the interface between price-quantity decisions and information acquisition. For example, in the standard theory of pricing and production under uncertainty, the firm is assumed to maximize its profit for a given distribution of product demand (or factor supply) shocks. In the New Classical Macroeconomics, agents formulate their price expectations and make their production and employment decisions on the basis of predetermined information sets. In the conventional literature on job search, (for example, Dale Mortensen, 1970; Edmund Phelps, 1970), workers make random draws from a wage distribution and, if they have imperfect information about the distribution, they may use their wage observations to update their estimates of the distribution.

Learning from experience involves more than acquiring information by taking random samples. Rather it involves agents' use of their price-quantity decisions to determine what sort of information they will receive. In example B, when a worker sets his wage at \$20 per hour (and then observes whether this wage was too high or too low), he acquires a different information set than when he sets his wage at \$30 per hour.

The existing literature on how agents find out about the demands they face is surprisingly small (for example, Michael Rothschild, 1974); the literature on how price-quantity decisions are used to reveal information is smaller yet (for example, Edward Lazear, 1986).

\*Department of Mathematics, London School of Economics, Houghton Street, London WC2A 2AE, and Department of Economics, Birkbeck College, University of London, 7 Gresse Street, London W1P 1PA, respectively.

### I. Learning from Experience

Learning from experience involves three interrelated economic activities: 1) making price-quantity decisions under uncertainty on the basis of current information; 2) observing the transactions generated by these decisions (i.e., "gaining experience"); and 3) on the basis of these observations, making inferences about the uncertain market conditions ("learning").

In abstract terms, a price or quantity decision  $g$  may be interpreted as an "experiment" which determines whether the actual "state of nature"  $z$ , lying in the uncertainty interval  $Z$ , belongs to a particular set. In particular, the decision  $g$  partitions the uncertainty interval  $Z$  and leads to an observation about which element of the partition contains the actual state of nature  $z$ . The search for the state  $z$  is carried out by a succession of price-quantity decisions, implying a succession of partitions and observations.

For instance, in example B, the worker's wage-setting decision is designed to determine whether his marginal product exceeds or falls short of his wage. Given that he knows his marginal product to be stable through time and to fall within the uncertainty interval  $[0, 1]$ , his wage  $W$  ( $0 \leq W \leq 1$ ) partitions the uncertainty interval as follows: if he receives a job offer, then he infers that his marginal product must lie in the interval  $[W, 1]$ ; if he remains unemployed, he infers that his marginal product lies in the interval  $[0, W]$ . In the next round of wage setting, the worker sets his wage within the new uncertainty interval, makes a new employment observation which leads to another inference about the uncertainty interval, and so on.

Needless to say, all price-quantity decisions are not equally "informative." In example A, the firm first supplies  $S$  goods and then, after sales have taken place, observes the stock of inventories. The supply decision partitions the demand uncertainty interval into a finite interval and a continuum of singleton sets: if there is a stock-out, then demand must be at least as large as the quantity supplied; if the inventory stock is positive, then the level of demand may be

inferred precisely. Consequently, the greater is the firm's supply, the finer the resulting partition (i.e., the smaller is the uncertainty interval inferred from a stock-out and the larger is the set of demand levels that may be inferred precisely), and thus the more informative is the supply decision.

Agents make each price-quantity decision in the face of a tradeoff between their current expected welfare and the value of the information (in terms of future expected welfare) obtained through their decision. In the optimal decision, the agents acquire just enough information through time to maximize the present value of expected welfare.

### II. Production, Product Demand, and Inventories

The way in which learning from experience generates high-low information and the implications of this behavior for the time path of price-quantity decisions may be clarified by studying examples A and B in detail. We begin with example A.

To set the stage, consider the firm's production problem in a one-period context. The firm does not know the level of product demand  $D$  that it faces (at a given price). However, it has a prior notion of the density of  $D$ , that (for expositional simplicity) we assume is uniform in the interval  $[X, 1]$ , where  $X$  is a constant,  $0 \leq X \leq 1$ .

The firm first decides to put a quantity ( $S$ ) of goods up for sale and then observes how much of this quantity is sold. (The quantity  $S$  is equal to the quantity produced plus the stock of inventory carried forward from the past.) The firm's one-period objective is to set the supply  $S$  so as to maximize its profit or, equivalently, to minimize its opportunity cost.

We define this opportunity cost quite simply: Let the product price ( $p$ ), the unit cost of production ( $f$ ), and the unit cost ( $h$ ) of overproduction (due to depreciation and discounting) all be known positive constants. The probability of excess demand is  $\text{Prob}(D > S) = (1 - S)/(1 - X)$ . In that event, the firm bears an expected opportunity cost of  $E[p - f] \cdot (D - S) | D > S$  (i.e., the expected value of foregone profit, conditional on the

occurrence of excess demand). The probability of excess supply is  $\text{Prob}(S > D) = (S - X)/(1 - X)$ . Then the expected opportunity cost is  $E[h \cdot (S - D) | S > D]$  (i.e., the expected value of the cost of holding the unsold goods as inventory, conditional on the occurrence of excess supply).

The firm's one-period decision problem is to set its supply so as to minimize its total expected opportunity cost, given the demand uncertainty interval  $[X, 1]$ :

$$(1a) \quad \min_S C_{[X,1]} \\ = \text{Prob}(D > S) \cdot E[(D - S) | D > S] \\ + \text{Prob}(S > D) \cdot E[(S - D) | S > D],$$

where (for simplicity) we let  $(p - f) = h = 1$ . It is straightforward to show that the optimal supply and the associated opportunity cost is

$$(1b) \quad S_{[X,1]}^* = (1/2) + (1/2) \cdot X; \\ C_{[X,1]}^* = (1/4) \cdot (1 - X).$$

We now turn to a simple two-period model that illustrates how the firm's supply decision may be used as a learning tool. In our example, learning is possible only if the firm's current and future product demands are correlated, for otherwise the firm's current sales observation would reveal no new information about future demand. Accordingly, let us examine the role that learning from experience plays in the firm's supply decision by comparing the following two scenarios: In the *No-Learning Scenario*, the first- and second-period demands are not correlated; whereas in the *Learning Scenario*, the demand remains constant from one period to the next.

At the beginning of the first period, the firm is assumed to start with a demand uncertainty interval  $[0, 1]$  and puts a quantity  $S_1$  up for sale. At the end of that period, sales take place. The probability of a stock-out (zero inventory) in the first period is  $\text{Prob}(D > S_1) = (1 - S_1)$  and the associated opportunity cost is  $E[(D - S_1) | D \geq S_1]$ . The

probability of excess supply (positive inventory) in the first period is  $\text{Prob}(S_1 > D) = S_1$  and the associated opportunity cost is  $E[(S_1 - D) | S_1 > D]$ .

In the *No-Learning Scenario*, the firm's observation of its first-period inventory stock provides no information about the second-period demand, and thus the demand uncertainty interval remains  $[0, 1]$ . Consequently (by (1b)), the firm's second-period supply is  $S_2^* = 1/2$  and the second-period opportunity cost is  $C_{[0,1]}^* = 1/4$ .

Let  $Z(NLS)$  be the present value of the expected opportunity cost in the *No-Learning Scenario* and let  $\delta$  be the time discount factor. Then the firm's decision problem in the first period is to set supply  $S_1$  so as to minimize  $Z(NLS)$ , given that the second-period supply is optimal ( $S_2 = S_2^*$ ):

$$(2a) \quad \min_{S_1} Z(NLS) \\ = \text{Prob}(D \geq S_1) \\ \cdot \{ E[(D - S_1) | D > S_1] + \delta \cdot C_{[0,1]}^* \} \\ + \text{Prob}(D < S_1) \\ \cdot \{ E[(S_1 - D) | D < S_1] + \delta \cdot C_{[0,1]}^* \}.$$

It is straightforward to show (by (1b) and (2a)) that the optimal supply levels in the two periods are

$$(2b) \quad S_1^*(NLS) = S_2^*(NLS) = 1/2.$$

In the *Learning Scenario*, the information acquired by the firm depends on its first-period inventory stock:

(i) If there is a stock-out ( $D \geq S_1$ ), the firm infers that demand (in both periods) lies in the interval  $[S_1, 1]$ . Thus (by (1b)), it sets its second-period supply at  $S_2^* = (1/2) + (1/2) \cdot S_1$  and the second-period opportunity cost is  $C_{[S_1,1]}^*$ .

(ii) If the inventory stock is positive ( $D < S_1$ ), the firm can infer the precise level of demand in both periods. (In particular, demand is the difference between the quantity put up for sale and the inventory stock left



over.) Thus, in the second period, the firm sets its supply equal to this demand, so that the associated second-period opportunity cost is zero.

Let  $Z(LS)$  be the present value of the expected opportunity cost in the Learning Scenario. Then the firm's first-period supply problem in the Learning Scenario is

$$(3a) \quad \min_{S_1} Z(LS) = \text{Prob}(D \geq S_1) \cdot \{ E[(D - S_1) | D > S_1] + \delta \cdot C_{[S_1, 1]}^* \} \\ + \text{Prob}(D < S_1) \cdot \{ E[(S_1 - D) | D < S_1] + 0 \}.$$

Now the optimal supply in the first period is

$$(3b) \quad S_1^*(LS) = (\delta + 2)/(\delta + 4).$$

If there is a stock-out at the end of the first period, then the optimal supply in the second period is (by (3b) and (1b))

$$(3c) \quad S_2^*(LS) = (\delta + 3)/(\delta + 4);$$

otherwise,  $S_2^*$  is set equal to demand.

Observe that if the firm is myopic (i.e.,  $\delta = 0$ ), then the first-period production is the same in both scenarios. The reason is that when the future is unimportant to the firm, its production decision will not be made with a view to learning about product demand.

If the firm is not myopic (i.e.,  $\delta > 0$ ), then its first-period production is greater in the Learning Scenario than in the No-Learning Scenario. Intuitively, it is clear why this is so. The greater the firm's supply in the first period, the more information it gains about product demand (because there is a contraction of the uncertainty interval associated with a stock-out and a rise in the probability of inferring demand precisely). Note that, in the Learning Scenario, an increase in the discount factor  $\delta$  leads to a rise in production and a higher expected inventory stock.

The simple model above can be extended in a wide variety of ways. Through the ap-

plication of the mathematical theory of high-low search, it has been extended to cover infinite time horizons (Phillipe Aghion, Patrick Bolton, and Bruno Jullien, 1987; our paper, 1987a), "conservative" (minimax) production strategies rather than Bayesian updating (our paper, 1987b), simultaneous production and pricing decisions (our papers, 1987c, d), and random variations in actual product demand (Diane Reyniers, 1987).

### III. Wages, Productivity, and Unemployment

Now we turn to example B. A worker's productivity at a firm depends on both the characteristics of the worker and the characteristics of his or her job slot in the firm. Whereas the worker may be expected to know more about his ability and work effort than the firm does, the firm generally has an informational advantage regarding its technologies, factor supplies and product demands. In example B, we focus on the job slot characteristics as a determinant of worker productivity, and examine how a worker may gain information about his productivity through his wage-setting decision. The worker's job is assumed to be "idiosyncratic," so that he cannot gain productivity information by drawing on the experience of other workers.

We begin with a simple, one-period characterization of the worker's wage-setting problem. The worker has imperfect information about how many units of output  $Q$  he could produce at a particular job. Let the worker's prior density of  $Q$  be uniform in the interval  $[X, 1]$ , where  $0 \leq X \leq 1$ . (We assume that this prior corresponds to the density of the worker's actual output across all the available jobs.)

The worker first sets his wage  $W$ , and then observes whether he gains employment. The firm knows the worker's actual productivity ( $\hat{Q}$  units of output), and offers him a job if  $\hat{Q} \geq W$ , but not if  $\hat{Q} < W$ . If the worker gets the job, he receives the wage he asked for and then his utility is  $W$ ; otherwise, he is unemployed, with a utility which (for simplicity) is given as  $-1$ .

In this one-period context, the worker's decision problem is to set his wage so as to

maximize his expected utility, given the productivity uncertainty interval  $[X, 1]$ :

$$(4a) \quad \text{Max}_W Y_{[X,1]} = \text{Prob}(Q \geq W) \cdot W \\ + \text{Prob}(Q < W) \cdot (-1),$$

where the worker's prior probability of getting the job is  $\text{Prob}(Q \geq W) = (1 - W)/(1 - X)$ , and the prior probability of remaining unemployed is  $\text{Prob}(Q < W) = (W - X)/(1 - X)$ . It is easy to confirm that the optimal wage and expected utility are

$$(4b) \quad W^* = \max(0, X) = X; \quad Y_{[X,1]}^* = X.$$

Now consider a simple two-period model that illustrates how the opportunity to learn about productivity may affect the worker's wage decision. For this purpose, we once again compare two scenarios: In the No-Learning Scenario, the worker's productivity in the first period is not correlated with that in the second period; whereas in the Learning Scenario, the worker's productivity at his job slot remains constant through both periods.

At the beginning of the first period, the worker has a productivity uncertainty interval which we normalize to  $[0, 1]$  and sets his first-period wage  $W_1$ . At the end of the period, he observes whether he receives a job offer. The probability of receiving a job offer in the first period is  $\text{Prob}(Q \geq W_1) = (1 - W_1)$  and the associated first-period utility is  $W_1$ . The probability of no job offer in the first period is  $\text{Prob}(Q < W_1) = W_1$  and the associated first-period utility is  $-1$ .

In the No-Learning Scenario, the worker gains no new information in the first period and thus his productivity uncertainty interval in the second period remains  $[0, 1]$ . Consequently (by (4b)), his optimal second-period wage is  $W_2^* = 0$  and his second-period utility is  $Y_{[0,1]}^* = 0$ . (Recall that wages and utility are measured on the same scale as productivity, which has been normalized to  $[0, 1]$ .)

Let  $V(NLS)$  be the present value of expected utility in the No-Learning Scenario. Then the worker's first-period decision prob-

lem is to set his wage  $W_1$  so as to maximize  $V(NLS)$ , given that the second-period wage is optimal ( $W_2 = W_2^*$ ):

$$(5a) \quad \text{Max}_{W_1} V(NLS) \\ = \text{Prob}(Q \geq W_1) \cdot (W_1 + \delta \cdot Y_{[0,1]}^*) \\ + \text{Prob}(Q < W_1) \cdot (-1 + \delta \cdot Y_{[0,1]}^*).$$

By (4b) and (5a) it can be shown that the optimal wage levels in the two periods are

$$(5b) \quad W^*(NLS)_1 = W^*(NLS)_2 = 0.$$

In the Learning Scenario, the employment observation in the first period provides information about productivity in the second period:

(i) If the firm makes no job offer in the first period, then the worker turns to another firm in the second period and offers to work for the wage  $\tilde{W}_2$  (where the tilde denotes "another" firm). His productivity uncertainty interval at the other firm is  $[0, 1]$ . Thus (by (4b)), the wage is  $\tilde{W}_2 = 0$ , and the associated utility is  $Y_{[0,1]}^* = 0$ .

(ii) If the firm makes a job offer in the first period, then the worker applies for the same job slot in the second period. He infers that his productivity lies in the interval  $[W_1, 1]$ . Thus (by (4b)), his wage in the second period is  $W_2^* = W_1$  and the second-period utility is  $Y_{[W_1,1]}^* = W_1$ .

Let  $V(LS)$  be the present value of expected utility in the Learning Scenario. Then the worker's problem in the first period is to set his wage  $W_1$  so as to maximize  $V(LS)$ , given that his second-period wage is optimal ( $W_2 = W_2^*$ ):

$$(6a) \quad \text{Max}_{W_1} V(LS) \\ = \text{Prob}(Q < W_1) \cdot (-1 + \delta \cdot Y_{[0,1]}^*) \\ + \text{Prob}(Q \geq W_1) \cdot (W_1 + \delta \cdot Y_{[W_1,1]}^*).$$

Now the optimal wage levels in the two

periods are

$$(6b) \quad W^*(LS)_1 = W^*(LS)_2 \\ = \delta / (2 \cdot (1 + \delta)), \quad \tilde{W}(LS)_2 = 0.$$

As we can see, if the worker is myopic (i.e.,  $\delta = 0$ ), then the wage (in both periods) is the same in both scenarios. Clearly, if the future is unimportant to the worker, then he will make no attempt to learn.

Yet if the worker is not myopic (i.e.,  $\delta > 0$ ), then his wages will be set higher when there are opportunities to learn about productivity than when such opportunities are absent. To see this, observe that the worker's costs and benefits of a marginal increase in  $W_1$  are the same in both scenarios, with one exception: in the Learning Scenario, the worker realizes that the higher the wage at which he receives a job offer in the first period, the higher the wage he can safely achieve in the second period. Thus, the marginal benefit of a first-period wage increase (at any given  $W_1$ ) is greater in the Learning Scenario than in the No-Learning Scenario. As result, the worker sets his first-period wage higher in the former scenario and, if he then receives a job offer, his second-period wage will be higher as well.

Now consider what effect this learning from experience has on unemployment. Imagine an economy with two overlapping generations of workers: in any period, there are  $n$  identical "young" workers (who set their wage  $W_1$  so as to maximize their expected utility over this period and the next) and  $n$  identical "old" workers (who set their wage  $W_2$  so as to maximize their current utility). A worker's productivity varies from one job slot to another; in particular, the density of a worker's potential output across all job slots is assumed to be uniform over the interval  $[0, 1]$ . (For simplicity, we assume that the number of potential job slots is sufficiently large so that the employment decisions in one period have no significant effect on the density of productivities in the next.) In any time period, each worker finds one vacancy and makes a wage offer for that job slot. Each old worker who has been

employed in the previous period applies for the same job in the current period. All other workers choose job slots at random.

It is easy to show that the expected level of unemployment for the old workers in both scenarios is zero. By contract, the level of "youth unemployment" is higher in the Learning Scenario than in the No-Learning Scenario. In both scenarios, each young worker's uncertainty interval is  $[0, 1]$  and his expected unemployment level is  $\text{Prob}(Q < W_1) = W_1$ . This means that in the No-Learning Scenario (where  $W_1^* = 0$ ), there is no youth unemployment; however, in the Learning Scenario (where  $W_1^* = \delta / (2 \cdot (1 + \delta))$ ), the level of youth unemployment is positive:  $n \cdot \delta / (2 \cdot (1 + \delta))$ .

This example suggests that when workers have some market power in wage determination and when they anticipate that the outcomes of their wage decisions will reveal information about their productivities, unemployment may arise.

#### IV. Concluding Remarks

The message of this paper extends well beyond the two examples given above. Quite generally, we have been concerned with how agents find out about the demands and supplies they face. Our point of departure is the simple observation that buyers and sellers often gain information by observing the outcomes of their price-quantity decisions. There are, of course, many other ways of gaining information: customer surveys, news reports, chats with colleagues, and so on. However, learning from experience appears to be sufficiently pervasive to merit individual attention. This learning activity has an impact on agents' price-quantity decisions because they realize that different decisions reveal different sets of high-low information.

#### REFERENCES

- Aghion, Philippe, Bolton, Patrick and Jullien, Bruno, "Learning through Price Experimentation by a Monopolist facing Unknown Demand," Working Paper 8748, University of California-Berkeley, 1987.

- Alpern, Steve, and Snower, Dennis J., (1987a) "Inventories as an Information-Gathering Device," ICERD Discussion Paper No. 87/151, London School of Economics, 1987.
- \_\_\_\_\_ and \_\_\_\_\_, (1987b) "Production Decisions under Demand Uncertainty: the High-Low Search Approach," Discussion Paper No. 223, Centre for Economic Policy Research (CEPR), 1987.
- \_\_\_\_\_ and \_\_\_\_\_, (1987c) "A Search Model of Optimal Pricing and Production," Discussion Paper No. 224, CEPR, 1987.
- \_\_\_\_\_ and \_\_\_\_\_, (1987d) "Price-Quantity Decisions as Learning Instruments," mimeo., 1987.
- Lazear, Edward P., "Retail Pricing and Clearance Sales," *American Economic Review*, March 1986, 76, 14-32.
- Mortensen, Dale T., "A Theory of Wage and Employment Dynamics," in E. S. Phelps et al. eds., *Microeconomic Foundations of Employment and Inflation Theory*, New York: Norton, 1970.
- Phelps, Edmund S., "Money Wage Dynamics and Labor Market Equilibrium," in his et al. eds., *Microeconomic Foundations of Employment and Inflation Theory*, New York: Norton, 1970.
- Reyniers, Diane, "Active Learning about the Demand Distribution in the Newsboy Problem," mimeo., 1987.
- Rothschild, Michael, "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, October 1974, 9, 185-202.

# The Search Equilibrium Approach to Fluctuations in Employment

By CHRISTOPHER A. PISSARIDES\*

Search theory was applied originally by macroeconomists to the explanation of employment fluctuations following aggregate shocks. It was offered as an alternative to the Phillips-Lipsey view of the relation between wage inflation and unemployment, which was effectively criticized by Phelps and Friedman as lacking theoretical foundations. The models of Edmund Phelps (1970) and Dale Mortensen (1970) are the most notable examples in this phase of the theory's development.

Several criticisms were made of these models, focusing especially on their incomplete description of equilibrium and their failure to account for layoffs and job-to-job changes. Although subsequent research responded to these criticisms, the bulk of subsequent research in search theory focused on questions of labor supply, steady-state equilibrium and efficiency and not on the original theme of employment fluctuations. Mortensen (1986) offers a good survey of the main body of this research.

More recently, search theory has been re-applied to the question of employment fluctuations. My main purpose in this paper is to evaluate some of these more recent contributions. Have they overcome the criticisms of early search theory? Are they plausible as models of the business cycle? Is there any evidence to support them?

It is useful when discussing these models to make the distinction between driving forces (or sources of impulses) and economic (or propagation) mechanisms. Search theory is an economic mechanism that can be combined into a theory of employment fluctuations with any of the recent literature's three predominant driving forces, misperceptions,

aggregate real shocks and sectoral shifts. David Lilien and Robert Hall (1986) discuss briefly the economic mechanism in the models of Phelps and Mortensen but reject it, largely for the same reasons that search theory was rejected as a model of fluctuations in the early to mid-1970's (see in particular p. 1023 of their survey). I argue here that this critique is unjustified in the light of more recent work in the area. I distinguish between two approaches in recent work, the reservation-wage approach and the trade-frictions approach.

## I. The Reservation-Wage Approach

The economic mechanism in the models of Phelps and Mortensen (1970) is job acceptance by searching workers and job quits by employed workers. A useful recent formalization of a model along these lines, with rational misperceptions, is contained in Randall Wright (1985). I call this the reservation-wage approach to employment fluctuations because the critical decision variable in the economic mechanism is the reservation wage.

In Wright's model, a worker samples from a distribution of wage offers. Each wage is the product of a job-specific productivity component and a general component, say price. When sampling, the worker cannot distinguish between the two components so he or she solves a Lucas-type signal extraction problem, that leads to a reservation-wage rule for the product of the two components. One period later he learns the true "real" component and applies a new reservation-wage rule to it. Since the two reservation wages do not generally coincide, some quits take place after the information is revealed.

The model exhibits persistence that, because of the learning mechanism, extends beyond the first order. However, the combi-

\*Centre for Labour Economics, London School of Economics, Houghton Street, London WC2A 2AE.

nation of misperceptions and reservation-wage rules limits the generality of this approach to fluctuations.

The empirical limitations of misperceptions models have been widely discussed. As a driving force, misperceptions are plausible only if the cost of the mistaken perceptions to the individual is very small, otherwise individuals will have an incentive to obtain up-to-date information before making decisions. In the case of search models, the cost of misperceptions is associated with making the wrong decision about one's job, which is anything but small.

Similar criticisms can be made of the use of reservation-wage rules. The reason for the persistence in unemployment under these rules is imperfect information about the location of the better-paying (and higher productivity) jobs. This is not likely to last for long when the individual's costs from the imperfect information are as high as they are in search models.

A more serious drawback of the reservation-wage approach is that although there is imperfect information about the location of jobs, the number of locations where jobs can be found does not vary over the cycle. The models must exhibit shortage of jobs of some kind, otherwise there would be no need for search. By emphasizing the reservation-wage aspects of search, these models implicitly assume that job availability is unchanging during the cycle, which contradicts the available evidence (James Medoff, 1983; my paper, 1986). By contrast, the second approach to search, discussed below, emphasizes variations in job availability over the cycle. Before turning to it, I discuss briefly one of the frequent criticisms made of search models: their inability to explain cyclical fluctuations in quits and layoffs.

## II. Job Separations

The failure of search models to explain the cyclical component of job separations is often cited as a major drawback of the search approach to employment fluctuations (see, for example, Lilien and Hall). This criticism, however, is valid only if two other conditions are met. First, fluctuations in employment

must be driven by fluctuations in job separations and not by fluctuations in accessions. Second, the fluctuations in job separations have to be something other than a simple equilibrium response to external shocks.

The first requirement is empirical. In Britain and the rest of Europe, fluctuations in the job separation rate contribute only a small fraction to fluctuations in the employment rate. Most of the fluctuations are accounted for by fluctuations in the rate at which workers enter employment (see my 1986 paper). In the United States, the job separation rate is more important than in Europe, but the unemployment exit rate is still an important contributor to fluctuations.

The second requirement, however, is more controversial. Search theory cannot explain "disequilibrium" job separations, in the sense of one side in the bargain, say the firm, causing a separation unwanted by the other side. But if, say, a job is hit by a shock that reduces the joint returns of employer and employee below the sum of their best alternatives, a separation takes place that is consistent with the optimizing behavior described by search theory.

Recent search-theoretic models of employment fluctuations rely on exogenous job-specific shocks to explain the transition from employment to unemployment. Thus empirically, unless exogenous shocks are cyclical, the employment fluctuations that can be rationalized by the theory are not the ones that can be attributed to cyclical layoffs. Quits are not cyclical and a big part of them take place near the beginning of a job tenure. Search models can easily be extended to have this implication, but the contribution of this extension to the explanation of fluctuations is not likely to be important. For a complete theory of the cycle, search theory needs to be combined with a theory of layoffs and dismissals, based for example on the theory of contracts.

## III. The Trade-Frictions Approach

Empirical evidence shows that firms' recruitment efforts vary procyclically. The Beveridge curve, that shows the relation be-

tween unemployment and job vacancies over the cycle, has a strong negative slope. In the United States, the number of Help Wanted ads shows the same cyclical behavior as the number of job vacancies in Europe. The second strand in the recent search-equilibrium literature emphasizes the trade frictions that exist in product and labor markets, and the role of fluctuations in job availability over the cycle.

The mechanism of this approach relies on the "search externalities" that arise because (a) during a short time interval, searching firms and workers make job contacts with probability less than one, and (b) the rate at which they make job contacts is a function of the number of traders in the market. The interaction between reservation wages and the wage-offer distribution is not important for these models, so I restrict myself to a discussion of equilibrium with a degenerate wage distribution.

A simple formalization of the central idea in this approach is the following. Suppose there are  $u$  searching workers and  $f$  searching firms. The typical worker puts into search  $c$  "efficiency units" of effort (his or her search intensity) and the typical hiring firm puts into it  $a$  efficiency units of effort (its advertising intensity). Job contacts are made according to the "search technology"  $x(cu, af)$ , that has positive first-order partial derivatives and elasticity with respect to each of its arguments less than 1. A worker supplying one efficiency unit during search makes a job contact at the rate  $x(cu, af)/cu$ . The number of efficiency units supplied by each worker is endogenous and determined by utility maximization. If the number supplied by worker  $i$  is  $c_i$ , the process of job arrival is Poisson with rate  $p_i = c_i x(cu, af)/cu$ . Similarly, a firm  $j$  supplies  $a_j$  units and makes a contact at the rate  $q_j = a_j x(cu, af)/af$ . The cost of  $c_i$  search units to the worker is a time cost, but the firm may use its labor force to search for new hires, so the cost of  $a_j$  is a labor cost (for example, it may be proportional to wages). In equilibrium all workers choose the same search intensity and if all firms are of the same size they all choose the same advertising intensity. The equilibrium  $c$  and  $a$  de-

pend, among other things, on each other and on  $u$  and  $f$ .

Empirically,  $af$  is a proxy for the number of job vacancies, or for the column-inches in the Help Wanted index.  $u$  is a proxy for unemployment and it varies according to  $\dot{u} = s(1 - u) - pu$ , where  $s$  is the frequency of job-specific shocks that lead to separations and  $p$  is equilibrium  $p_i$ , evaluated at  $c_i = c$ . Thus, unemployment exhibits persistence.

By the assumptions made on the search technology,  $p_i$  depends negatively on the inputs of other workers ( $cu$ ) and positively on the inputs of firms ( $af$ ). The former is a congestion externality and the latter an external economy; in general, firms and workers will ignore the effects of their choices on the transition probabilities of other firms and workers and equilibrium will be inefficient. Similar externalities affect the firm's hiring rate  $q_j$ .

These externalities provide the crucial mechanism that propagates shocks in this class of models. Because of the externalities, the models imply that the returns from trade depend on the tightness or thinness of the market. Search is not necessary for this property to hold and stylized macro models without search that satisfy this property have been developed (for example, by Peter Howitt, 1985). But search is an obvious rationalization of such mechanisms.

#### IV. Multiple Equilibria?

The existence of the external trade economy makes the equilibrium trade activity of workers a rising function of the activity of firms, and vice versa. This property of equilibrium underlies Peter Diamond's (1982) "coconut" model, where it is claimed that plausible restrictions on the search technology can lead to more than one equilibrium. At a low-level equilibrium, not many firms hire and workers come less frequently into the market to look for jobs. At a high-level equilibrium, more agents come into the market and the return from trade is higher. More of it is demanded and equilibrium unemployment rates are lower. Howitt's (1985) model is also characterized by multi-

ple equilibria in some cases, for similar reasons.

Multiplicity of equilibrium as an economic mechanism is radically different from any of the existing models of fluctuations because it can exacerbate the effects of shocks to a far greater extent. For example, even minor misperceptions of nominal shocks may have long-lasting effects if they are capable of shifting the economy from one equilibrium to another. The adjustment dynamics of models with multiple equilibria, however, have not been explored and local stability may be an important issue in the empirical relevance of the models. Empirically, the only relevant question at the current stage of development is whether the conditions needed for multiple equilibria can stand up to the evidence.

A necessary, though not sufficient, condition for multiplicity is that the search technology  $x(cu, af)$  should have increasing returns to scale; that is, the external economy should dominate the congestion externality in the equilibrium transition rates (see Diamond, 1984; Howitt, 1985). I tested this condition with British data on unemployment and job vacancies and found evidence against it (see my 1986 paper). I found the external economy to be strong and significant on both firms and workers: the probability of leaving unemployment depends positively on the vacancy rate and the vacancy rate depends positively on the unemployment rate. But I also found strong evidence for constant returns to scale in the search technology, with the elasticity with respect to unemployment estimated at 0.7 and the elasticity with respect to job vacancies estimated at 0.3.

### V. Fluctuations with Unique Equilibrium

If search equilibrium with trade frictions is unique, the propagating mechanism works mainly through the hiring activities of firms, that is, through job availability. Thus the models have some "Keynesian" features, despite their equilibrium nature. Models of fluctuations with this property were developed by myself (1985, 1987) and by Howitt (forthcoming).

In both my papers, the driving force is an aggregate real shock of the labor-augmenting type. Perfectly anticipated permanent productivity shocks do not influence the rate of unemployment: as in the neoclassical model of growth these shocks are fully absorbed by wages. But temporary shocks (or permanent shocks perceived at first as temporary) are not fully reflected in wages, so they feed into profits. If the temporary shock is positive firms come into the market with more jobs to take advantage of the higher profit and this raises trade activity at given search intensity. The external economy generated by the new jobs also raises intensity, leading to even more trade activity and to a rapid fall in unemployment. Unemployment returns to its steady-state level when the temporary shock ends (or when full information about the permanence of the shock is revealed) and profits return to normal.

The behavior of wages in these models is critical in the transmission of temporary shocks to employment. In my papers (1985, 1987), wages are determined endogenously by the meeting firm and worker, according to the Nash solution to the bargaining problem. Some kind of bargaining solution for wages is required when there are trade frictions, because each job enjoys some local monopoly power which shows up as economic rent. The rent is equal to the imputed value of the cost of trade which, as far as a matched pair is concerned, is sunk. In the Nash solution, whose use in this model can be defended by the argument of Kenneth Binmore, Ariel Rubinstein and Asher Wolinsky (1986), labor's alternative return is a strong influence on wages. In my paper (1987), I argued that labor's alternative return is its reservation wage, which is a forward-looking variable. When a shock is perceived as temporary, future wages are not expected to rise, so the reservation wage does not respond much to it. The Nash bargain tilts the division of the returns from the shock in favor of profits.

Thus reservation wages need not be passive in the trade-frictions model. Their role, however, is different from the one in the reservation-wage model. In the latter, a temporary real shock raises employment by rais-



ing actual wages by more than reservation wages. In the former, reservation wages hold down actual wages. Employment rises because the unresponsiveness of wages brings more jobs into the market.

The economic mechanism in Howitt's (1987) model also works through job availability, though the monopoly rents enjoyed by jobs do not play an important role in his model (and thus fully perceived real shocks have no effect on employment, unless they are accompanied by intersectoral shifts in productivities). Wages are simply proportional to marginal product. The driving force in the model is misperceptions: there are stochastic disturbances to relative and absolute prices and a Lucas-type signal extraction problem. When a firm experiences a rise in its relative price (actual or perceived), the return from hiring goes up, so the firm increases its hiring activities. If the rise in relative price is perceived but not actual, due to an increase in aggregate demand, other firms experience a similar rise and so the number of job vacancies rises, producing the business cycle response studied by Howitt. He shows that the behavior of employment can be described by a Lucas supply function, but because the mechanism in his model relies on trade externalities the full-information level of employment is not necessarily efficient.

Models with trade frictions can also be used to study the effects of intersectoral shifts in demand, like those studied by Robert Lucas and Edward Prescott (1974) and Lilien (1982). In the Lucas-Prescott model, where there are no trade frictions, an intersectoral shift of demand raises wages in one sector and reduces them in another. Labor leaves the depressed sector, but has to wait for a fixed period before entering the booming sector. Thus unemployment rises temporarily. In the trade-frictions model (Howitt's model is particularly well-suited to the study of this kind of shock), firms' hiring activities rise in one sector and fall in another, so average unemployment rises if there is a nonlinear response of job-matching rates to changes in trade activity, which is plausible. Of course, if intersectoral movement is more costly than intrasectoral movement, as in the

Lucas-Prescott model, unemployment may persist longer.

The trade-frictions model implies a strongly nonlinear response of unemployment to intersectoral shifts, which has not been explored by the empirical literature. This is because of the local monopoly power enjoyed by matched job-worker pairs. Small negative shocks do not induce job separations because the economic rents provide a cushion against them. Downward employment adjustment takes place through reduced hirings so it is slow. But if the intersectoral shifts are big enough to eliminate the economic rents of many jobs in the adversely affected sectors, there is a fast rise in unemployment through job separations that is matched by a much slower employment rise in the expanding sectors.

## VI. Conclusions

The search equilibrium models with trade frictions provide a framework for the study of fluctuations with many promising features. They have an explicit theory of unemployment where the private gains from trade are fully exploited. They are especially well-suited to the study of dynamics out of the steady state. The adjustment path and the final equilibrium depend on both supply-of-labor factors through search intensity and demand-for-labor factors through job availability and hiring intensity. Each job enjoys local monopoly power, so wage determination can be analyzed in terms of a bargaining process, with the many possibilities that this opens up for noncompetitive wage behavior. The mechanism that transmits shocks to employment has the Keynesian feature of limited job availability: a positive nonneutral shock is followed by a relaxation in the number of jobs offered and by a gradual rise in employment. Most of these features of the models have only just begun to be explored.

## REFERENCES

- Binmore, Kenneth, Rubinstein, Ariel and Wolinsky, Asher, "The Nash Bargaining Solution in Economic Modelling," *Rand Journal of Economics*, Summer 1986, 17, 176-88.

- Diamond, Peter A.**, "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, October 1982, 90, 881-94.
- \_\_\_\_\_, *A Search Equilibrium Approach to the Micro Foundations of Macroeconomics*, Cambridge: MIT Press, 1984.
- Howitt, Peter**, "Transaction Costs in the Theory of Unemployment," *American Economic Review*, March 1985, 75, 88-100.
- \_\_\_\_\_, "Business Cycles with Costly Search and Recruiting," in *Quarterly Journal of Economics*, forthcoming.
- Lilien, David M.**, "Sectoral Shifts and Cyclical Unemployment," *Journal of Political Economy*, August 1982, 90, 777-93.
- \_\_\_\_\_, and **Hall, Robert E.**, "Cyclical Fluctuations in the Labor Market," in O. C. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, Vol. 2, Amsterdam: North-Holland, 1986.
- Lucas, Robert E. and Prescott, Edward C.**, "Equilibrium Search and Unemployment," *Journal of Economic Theory*, February 1974, 7, 188-209.
- Medoff, James L.**, "U. S. Labor Markets: Imbalance, Wage Growth and Productivity in the 1970s," *Brookings Papers on Economic Activity*, 1:1983, 87-120.
- Mortensen, Dale T.**, "A Theory of Wage and Employment Dynamics," in E. S. Phelps et al. eds., *Microeconomic Foundations of Employment and Inflation Theory*, New York: W. W. Norton, 1970.
- \_\_\_\_\_, "Job Search and Labor Market Analysis," in O. C. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, Vol. 2., Amsterdam: North-Holland, 1986.
- Phelps, Edmund S.**, "Money Wage Dynamics and Labor Market Equilibrium," in his et. al., eds. *Microeconomic Foundations of Employment and Inflation Theory*, New York: W. W. Norton, 1970.
- Pissarides, Christopher A.**, "Short-Run Equilibrium Dynamics of Unemployment, Vacancies, and Real Wages," *American Economic Review*, September 1985, 75, 676-90.
- \_\_\_\_\_, "Unemployment and Vacancies in Britain," *Economic Policy*, October 1986, 3, 499-559.
- \_\_\_\_\_, "Search, Wage Bargains and Cycles," *Review of Economic Studies*, July 1987, 54, 473-84.
- Wright, Randall**, "Job Search and Cyclical Unemployment," *Journal of Political Economy*, February 1985, 94, 38-55.

# Self-Fulfilling Optimism in a Trade-Friction Model of the Business Cycle

By ALLAN DRAZEN\*

Models of economic activity with frictions in coordinating trading have been shown to be capable of generating multiple steady states. (Peter Diamond, 1982, is the pioneering work; see my 1987a paper for a general discussion.) Less work has been done on out-of-steady-state dynamics in such models, which would enable us to examine what sort of fluctuations these models may generate. The absence of dynamics leaves open the question of which steady state the economy will reach, as well as whether the comovements of key variables resemble what is observed over the cycle.

In my earlier paper (1987b), I presented a model combining search and aggregate demand approaches to unemployment to show how spillovers between product and labor markets could yield multiple equilibria. Here, a highly simplified dynamic model based on his work is presented, in which (self-fulfilling) sales expectations determine to which steady state the economy converges. These expectations are summarized by the asset values of firms which are producing output relative to those that are not. Two types of dynamic paths leading to stationary solutions can arise. The first is a saddle path. In addition, for certain parameter values, stable limit cycles emerge. Interestingly, over this cycle, asset values of firms (which one could interpret as "stock market" values) lead economic activity.

## I. Model Setup

There is an equal, large fixed number  $L$  of homogeneous firms and workers. A firm consists of a single job with a fixed coefficient technology: a job (i.e., firm) employs one worker, one unit of perishable output being produced independent of the identity of the firm or worker. Firms may be filled or vacant; workers may be similarly employed or unemployed.

The key to fluctuations in this type of model is that unemployed workers and vacant firms must find one another in order for a vacant slot to be filled. Let contact be made according to a Poisson process, with instantaneous rate of contact  $\pi$ . Since all jobs and workers are identical, contact will always lead to employment, so that  $\pi$  will also be the rate at which firms fill jobs (or workers find employment). I assume that  $\pi$  is an increasing, concave function of the recruiting intensity  $\beta$  with which a vacant firm seeks workers.  $\pi$  is taken to be independent of other firms's recruiting efforts, and of the aggregate unemployment rate. (The second assumption is made to highlight a somewhat different source of multiplicity than that in Diamond. There the probability of meeting a trading partner was increasing in the fraction of the labor force seeking matches, and this aspect was crucial for multiple equilibria. I eliminate this to focus on spillovers across markets.) Flow recruiting costs are represented by an increasing, convex function,  $g(\beta)$ .

I assume that jobs are terminated at an exogenous rate  $\delta$ , where termination is also a Poisson process. Therefore, an infinitely lived firm may be in one of two states—filled and producing, or vacant and recruiting—with an instantaneous probability  $\delta$  of moving from the former to the latter state, and of  $\pi$  of moving from the latter to the former state.

\*Department of Economics, Tel-Aviv University, Ramat-Aviv, Israel, 69978, and University of Pennsylvania, Philadelphia, PA, 19104. I thank Costas Azariadis, Roger Farmer, and participants in informal seminars at the Institute for International Economic Studies, Stockholm, and the University of Pennsylvania for helpful comments. Financial support from the National Science Foundation, grant SES-8706808, is gratefully acknowledged.

Similarly, a worker may be either employed or unemployed, with the same transition probabilities between states.

We may now derive the value of a representative filled and vacant firm, denoted  $F$  and  $V$ , respectively, by noting that the value multiplied by the interest rate must equal the income flow in that state plus the expected capital gain or loss coming from a change in the value or from a change in state. Though the individual firm switches states stochastically, the value of a representative firm in either state is nonstochastic. (These equations may also be derived as the limit of a discrete time formulation.) I assume the interest rate is equal to the constant discount rate  $r$ .

For the vacant firm we have in continuous time

$$(1) \quad rV = -g(\beta) + \pi(\beta) \cdot (F - V) + \dot{V},$$

where a dot over a variable denotes time derivative.

For a filled firm, the demand curve it faces should be derived from a fully specified model of consumer behavior, where actual sales depend on the price the firm charges. Space limitations prevent spelling out such a model (which is presented in my earlier paper, 1987b). The key implication for this paper is that sales depend on aggregate market conditions. This can be summarized by assuming that a filled firm can sell a fraction  $c$  of its output in any period, where  $c$  is a decreasing function of the aggregate unemployment rate  $u$ .<sup>1</sup> Any output not sold

perishes. Profits in a given period are  $c(u) - w$ , where  $w$  is the wage rate in terms of output. It is assumed that the wage rate is determined as a splitting between firm and worker of the surplus associated with a match. For simplicity, let the split be such that the real wage is a fixed proportion of sales, so that firm profits will also be a decreasing function of  $u$ , which I denote  $\sigma(u)$ . (A more sophisticated rule would have this proportion, and hence  $\sigma$ , also dependent on expectations of future sales and hence  $H$  either directly or via  $g(\beta)$ . This turns out to change the model's dynamic behavior.) The value of the filled firm then obeys the equation

$$(2) \quad rF = \sigma(u) - \delta \cdot (F - V) + \dot{F}.$$

## II. A Simple Dynamic System

If we define the excess of the value of a filled over a vacant firm as  $H = F - V$ , we may subtract (1) from (2) and rearrange to obtain

$$(3) \quad \dot{H} = (r + \delta + \pi(\beta))H - \sigma(u) - g(\beta).$$

With an equal, fixed number of firms and workers, the unemployment rate equals the vacancy rate and evolves as the flow into unemployment minus the flow out of unemployment. We thus have

$$(4) \quad \dot{u} = \delta \cdot (1 - u) - \pi(\beta)u.$$

Optimal recruiting intensity  $\beta$  is determined by the vacant firm. Integrating (3) we obtain

$$(5) \quad H(t) = \int_{x=t}^{x=\infty} [\sigma(u(x)) - g(\beta(x))] \times e^{-(r+\delta+\pi(\beta(x)))(x-t)} dx.$$

Interpreting  $\beta$  in (5) as the optimal value of  $\beta$  for  $x \geq t$ , (5) gives the maximized value of  $H$  from  $t$  onwards. Using this interpretation,

<sup>1</sup>More specifically, consider a Dixit-Stiglitz model in which symmetric price-setting firms sell differentiated products and producing firms must incur a constant marginal cost per unit actually sold. Suppose for example that both the upper-tier utility function, over an aggregate of the differentiated product and a nonproduced good, and the subutility function over the differentiated product are CES. With income proportional to the number of firms producing, one may show that the equilibrium price of the differentiated product is a constant markup over marginal cost and that the share of consumption on differentiated products is increasing in the number of firms producing, which is  $(1-u)L$ . Endogenous sales per firm can then be shown to be decreasing in the aggregate unemployment rate.

we now integrate (1) to obtain

$$5) \quad V(t) = \int_{\tau=t}^{\tau=\infty} [-g(\beta(\tau)) + \pi(\beta(\tau))H(\tau)] e^{-r(\tau-t)} d\tau.$$

$V(t)$  has the interpretation of the value of the vacant firm on the assumption that  $\beta$  is chosen optimally from any  $\tau \geq t$  onwards. Optimal  $\beta(\tau)$  may then be found by pointwise maximization of (6) yielding

$$7) \quad -g_{\beta}(\beta(\tau)) + \pi_{\beta}(\beta(\tau)) \cdot H(\tau) = 0,$$

where subscripts denote partial derivatives. The solution to (7) may be written in implicit form as  $\beta(H)$ . The concavity assumptions above imply that recruiting intensity  $\beta$  increases with  $H$ . Substituting  $\beta(H)$  into (3) and (4) we then obtain a system of two differential equations in the two variables  $H$  and  $u$ .

### III. Out-of-Steady-State Behavior

The dynamics of the system may be represented by a phase diagram, as in Figure 1. One may show that the  $\dot{u} = 0$  and  $\dot{H} = 0$  loci are downward sloping in  $u$ - $H$  space, given the above assumptions about  $g(\beta)$ ,  $\pi(\beta)$ , and hence  $\beta(H)$ , but need be neither concave or convex and may well "wiggle." Hence even for simple specifications of the underlying functions, there can be multiple intersections.

Taking the linear approximation of the system around each intersection, one may show that intersections where  $\dot{u} = 0$  locus cuts the  $\dot{H} = 0$  locus from above are saddle points. (One considers the properties of the Jacobian matrix.) The existence of multiple saddle points means that a given unemployment rate may be consistent with more than one saddle path, and hence  $H$ . Remember that  $H$  has the interpretation of the (expected) excess value of the filled over the vacant firm. This excess depends on expectations of future sales, which is in turn determined by future expected unemployment. When a value of  $u$  is consistent with several values of  $H$ , optimism on the part of vacant

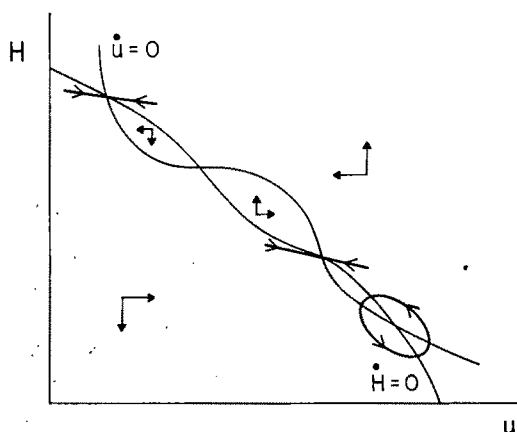


FIGURE 1

firm about future sales possibilities of filled firms will be reflected in a high value of  $H$  and a high saddle path, and the economy will converge to a low unemployment steady state. If vacant firms are more pessimistic about future sales,  $H$  will be lower, implying movement along a lower path, and the economy will converge to an equilibrium with higher unemployment. In each case, the expectations of economic activity are rational, given the actual future movement of the system, or, one may say, self-fulfilling.

At those intersections where the  $\dot{u} = 0$  locus cuts the  $\dot{H} = 0$  locus from below, the steady states are unstable for  $r \geq 0$ , which is the trace of the Jacobian matrix. Since stability at these intersections depends on whether the trace is positive or negative, the behavior of the system will change ("bifurcate") at  $r = 0$ . One may then show that the conditions of the Hopf bifurcation theorem are satisfied by this model for  $r = 0$  (see John Guckenheimer and Phillip Holmes, 1983, for a statement and discussion of the theorem, as well as for the stability test mentioned below), meaning that we get local limit cycles on one side of  $r = 0$ . If cycles occur for  $r > 0$ , they are economically meaningful and are stable. One may show that for certain functional forms and ranges of parameter values, the model can generate stable cycles. In short (and less formal terms), this model can generate a stable business

cycle of the sort illustrated in Figure 1 in the absence of exogenous stochastic shocks.

#### IV. Cyclical Behavior of Economic Aggregates

What is the behavior of economic variables over such a cycle? Especially interesting is that  $H$ , representing the asset value of producing firms relative to vacant firms, leads economic activity as represented by the cycle drawn in Figure 1. Beginning let's say with a "boom" period of rising  $H$  and falling  $u$  (a northwesterly movement),  $H$  will peak and begin to turn down before the downturn in  $u$ . Similarly, when  $u$  is rising and  $H$  is falling, the upturn in  $H$  will lead the upturn in  $u$ . In this model, movements in asset values "predict" economic activity.

If price setting were introduced into the model (see fn. 1) the behavior of prices would depend on how consumers' utility functions were specified, determining the nature of demand curves firms face. For the CES example in footnote 1, prices do not vary over the cycle. Somewhat more complicated specifications yield both prices and sales per firm (and hence per employed worker) being procyclical. In the simple model presented here, profits and real wages are procyclical, though this was very much built into the model. A

more complete characterization of the behavior of variables such as these depends on a fuller specification of how both prices and wages are set. The next step is to enrich the model in order to attempt to better replicate the behavior of variables over the cycle. Even the simple model, however, indicates that this approach can generate rich and not unrealistic dynamics, even when we consider only rational expectations paths.

#### REFERENCES

- Diamond, Peter A., "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, October 1982, 90, 881-94.
- Drazen, Allan, (1987a) "Reciprocal Externalities Models of Low Employment," *European Economic Review*, February/March 1987, 31, 436-43.
- , (1987b) "Involuntary Unemployment and Aggregate Demand Spillovers in an Optimal Search Model," mimeo., rev. 1987.
- Guckenheimer, John and Holmes, Phillip, *Non-linear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, New York: Springer-Verlag, 1983.

## "THE NEW INDUSTRIAL STATE" AFTER TWENTY YEARS

### Time and the New Industrial State

By JOHN KENNETH GALBRAITH\*

The controlling advice for this paper comes from Winston Churchill: "I've often," he said, "had to eat my words, and, in general, I have found them a wholesome diet." I will follow Churchill's example but with restraint. On returning to a book now twenty years in print, I will not deny myself the pleasure of affirming what I believe still to reflect the reality. Or of rebuking in a tolerant way those of our profession who do not accept what now seems to be evident. But I will not be wholly reluctant in admitting error or insufficient foresight. In this latter effort, I do not doubt that I will have the help of my distinguished fellow members of his panel.

A willing recognition, if not of error, at least of obsolescence, is, in fact, implicit in the view of economics that I avow in *The New Industrial State* (hereafter *TNIS*). I see economics as a subject in constant accommodation to social, political and institutional change and not, certainly, as a search for, and expression of, unchanging truth. It follows that I cannot suppose myself to be exempt from the effects of this insistent process of change.

Some compelling developments of the last twenty years have, I venture, served to confirm my earlier stated view of economic life. In *TNIS* I offered as a central, even indispensable, concept a bimodal view of the developed economic system: in the United States, as in the other industrial countries, a comparative handful of great corporations on the one hand, vast numbers of small entrepreneurial enterprises on the other. On the validity of this view of the economic

system I would, needless to say, still insist. And also on the marked differences in both structure and motivation as between the two parts of the system. One part conforms in greater or lesser degree to the norms of the textbook market. The other has a distinctly separate character and motivation. One yields in reasonable measure to the theoretical and mathematical refinements of modern microeconomics. The other, even as amended by monopoly and oligopoly theory, does not.

Specifically, in the sector of the large corporations authority passes from owners or capitalists to the bureaucratic apparatus of the firm that I called the technostructure. Given the diverse knowledge, experience, and personality that are required for decision, this movement is inevitable. In this structure, power along with pecuniary return becomes a major motivational goal. And the pursuit of pecuniary return has marked ambiguity. Management is presumed to seek maximization of such return not on its own behalf, but on that of passive and powerless stockholders.

The corporate part of the economy, I held also, exercises an influence or control over the larger economic, political, and social context that far exceeds anything available to the entrepreneurial firm, or that is envisioned in conventional oligopoly theory. Going beyond prices, it includes the ensuring and controlling of sources of supply, the associated cost management, the shaping of consumer response, and the obtaining of the relevant government action and support. This last I held to include support to aggregate demand and to the market for specific products, a notable case being that of the weapons industry. Surveying the extent and purpose of this control, I was led in later editions of *TNIS* to refer to the great corporate two-thirds of the bimodal system as the Planning

\*Department of Economics, Harvard University, Cambridge, MA 02138.

System. I came partly to this change, I confess, because of the discomfort I thought it would cause my more theologically orthodox readers, in the admittedly unlikely event of there being any.

As to the bimodal character of modern industrial organization, not alone in the United States but in the industrial economies in general, I would, as noted, stand firm. The entrepreneur, the economist's only hero, is much celebrated in our time; the hard statistics on industrial concentration are not to be escaped, as Wassily Leontief and many others have shown.

Neither now to be escaped is the conflict of interest between technostucture or management and ownership—what in *TNIS* I called The Approved Contradiction. In these last years we have seen this motivational contradiction in the modern corporation in compelling form. The corporate raider and the related acquisition and takeover mania have their source in the differing objectives of management and owners. So also do the present much-featured extravagances in executive compensation and, needless to say, in the thoughtfully arranged golden parachutes. This contradiction, foretold, I should be prompt to say, in the earlier work of Adolf Berle and Gardiner Means and of the recently deceased James Burnham, I take to be fully affirmed. However, unless my reading of current economic discussion is sadly deficient, I cannot think that it has yet made its way in any adequate fashion into modern microeconomic comment and instruction.

Also affirmed, I would urge, is the effort of the modern great enterprise to influence or control its context, including, notably, its consumer response. That the large corporation seeks to extend its authority on from price determination to the other factors bearing upon its operations is surely evident. The phraseology of this effort—cost management, strategic planning, marketing strategy, public relations, PACs—is entrenched in our everyday language. Only in professional economic analysis and instruction do we succeed in setting it aside.

Economic orthodoxy has been especially resistant where any impairment of the con-

cept of consumer sovereignty is involved. And understandably, if not rightly, so. Service to the sovereign will of the consumer is, after all, what gives social purpose to economic life—and to the study of economics. Much of the social justification of our subject is lost when the wants of the consumer are shown to be even partially the contrivance of those who serve them. Yet the evidence of producer influence on consumer response is hardly to be avoided. No effort is more visible; a huge industry is devoted thereto. Not even the cold eye of the economist can entirely escape the insistent world of television.

The frequent economic argument that this vast effort cancels itself out and leaves consumer sovereignty intact must, even as a defense of reputable orthodoxy, be counted more than ordinarily inadequate. Paul Samuelson and William Nordhaus, in their most recent edition (*Economics*, 12th ed., 1985), do make a bow to the diminishing pressure of wants and the consumer "response to pressures of fashion and advertising." But they quickly withdraw to a world of unmet public and private needs. Scarcity is essential, and the authors, not finding it sufficiently available in the modern industrial world, extend their reach to embrace the poverty-ridden lands of the Third World, to which, it cannot be supposed, their excellent work normally goes. Campbell McConnell stands firmly for consumer sovereignty on the strength of the public rejection of the new components of the Coca Cola. (See *Economics*, 10th ed., 1987.) For long, the decisive proof had been in the rejection of the Edsel.

There are other, rather broader matters urged in *TNIS* on which I would still insist—the diminishing role of the trade union in the modern industrial society; the service of conventional market theory as an ideologically convenient design for concealing the exercise of power by the great corporation, a point for later mention; the intellectual and social accommodation that makes the production of goods and services the supreme test of social achievement; our failure to miss seriously the goods when, in a recession, they are not produced but the grievous



ain in such a recession from missing the employment and income flowing from production; our differential and highly conditional attitudes as to work and leisure; the road tendency to a bureaucratic and cultural convergence of large capitalist and socialist enterprises. But I have more than sufficiently urged my earlier position; I must now turn to matters on which I have been overtaken by time or, on occasion, as I fondly hope, by deeper insight.

The first of the needed revisions concerns the control of context exercised by the great firms of the planning system. As regards the state, this has increased in important respects. In the United States as also in Canada, Britain, and elsewhere, conservative governments have brought a conjunction of state and corporate power that I did not envisage twenty years ago. It cannot be said, however, that this, especially in broad macroeconomic design, has reflected a wholly intelligent expression of longer-run corporate interest. My inclination then was to a much more rational view, one which accorded with the broad Keynesian view as regards both inflation and unemployment and underproduction. I did not foresee the surrender to the grave political asymmetry of Keynesian fiscal policy—to the superior political attraction of tax reduction and expenditure enhancement as compared with the reverse. Nor did I foresee the consequent escape into monetarism, including the attraction to affluent individuals and institutions of high real interest rates, a basic instrument of monetary policy. Like others, while doubting the efficacy of monetary policy, I assumed it to be socially and politically neutral. It assuredly is not. It greatly favors those with money to lend, and those with money to lend are usually richer than those not so endowed. Economists, I am persuaded, commit no greater error than in failing to recognize this.

There have also been important microeconomic developments that I did not foresee. In 1967, my view was, in some measure, of a closed America-dominated corporate structure extending its reach internationally by way of American multinational or transnational enterprises. I did not foresee the inva-

sive thrust into this structure by Japan and other countries. This, to put it mildly, has introduced a new element, substantially beyond the influence and control of the firms of the corporate or planning system.

However, I would still hold that the great corporate enterprises continue to function within a system of extensively known parameters, which allow, in turn, for the associated planning. Failure to recognize this is another major error of economics; it is one reason why those aspiring to a career in the real economic world study not economics but business administration. Although the parameters within which American business planning proceeds are, as noted, substantially less firm than they were twenty years ago, they are still the relevant circumstance. The modern corporate enterprise cannot, accordingly, be brought within the framework of conventional and mathematical market theory. The present solution, which is to adjust the reality to what can be accommodated to the equations, is not something we should applaud or approve.

The second revision I would now make concerns the technostructure. As to its authority in the great enterprise, there is not, as I've indicated, any cause for doubt. Nor as to its necessity. Decision in the modern business enterprise calls on the knowledge and experience of diverse specialists, diverse talent, and diverse experience. This knowledge is brought together for a quality of decision far superior to that of any one man or woman. On few matters are we so committed to error as in the special intelligence, ability, and authority we attribute to the individual who, for the moment, presides over this structure and is its public voice, and who, on the day of retirement, returns to a total and often well-deserved obscurity. The securities markets celebrate that departure with total indifference. The modern executive is a creature of organization, not the reverse.

In making these points in *TNIS*, however, I unduly emphasized the intelligence and efficiency of the technostructure. I did not sufficiently see its deteriorative tendencies. These include its relentless multiplication of personnel; nothing so denotes prestige as the

number of one's subordinates, and nothing so minimizes the pain of thought and action as having others to whom these can be delegated. The recurrent dismissals by the modern corporation of large numbers of staff, always "in the interest of efficiency," is ample evidence of this circumstance. Additionally, in all large organizations there is a powerful tendency for wise action to be what is most companionate with what is already being done, talent to be what most resembles those already there.

These bureaucratic tendencies have come much to our attention in these last years. Enough were almost certainly in evidence twenty years ago to have justified more mention than they were given. A certain sclerotic commitment to what was already being done was evident to those of us who dealt extensively with the older industries such as steel and coal in World War II. In this regard, *TNIS* reflects more than a little of the economic contentment, perhaps even euphoria, in the years of the 1950's and 1960's.

Further, were I now writing this book—admittedly a forbidding thought—I would specifically stress more strongly the now sadly evident policy error in the conventional separation of microeconomics from macroeconomics. The latter, the legacy of Keynes, held the overall performance of the economy to be less than optimal in normal equilibrium. Correction then became the accepted responsibility of the state and the central bank. On the other hand, microeconomic performance, monopoly and other competitive imperfections apart, was considered broadly optimal or, in any case, free from the need for generalized public action.

I do not hope to see microeconomics rescued from the recreational technicality into which, for many, it has fallen. Paralleling what Dr. Johnson said of making money, it is the least harmful of professional pursuits. I would, however, like to see a separate and growing concern for the public policy issues arising in microeconomic performance. These now rival, perhaps exceed, macroeconomic concerns in their social urgency. The failure as to housing—the great industrial default of capitalism—the problems of energy and

oil, of agriculture, of the aging industrial sector, and the difficulties arising from the competitive relationship between the older and younger industrial economies, all call for such attention. Likewise the stubborn persistence of youth and minority unemployment, which are both resistant to macroeconomic remedy. It is to these matters, as much as to conventional macroeconomic issues, that public policy needs now to be directed. None of them is corrected by even the wisest macroeconomic action.

Finally, were I now writing this book, I would, as will be evident, be more appreciative of the difficulty that our subject matter has in accepting change. And also of the reasons therefor. Part of this difficulty, to repeat, lies in the perception of their discipline by economists as a science of unchanging ultimate truths. Only this is thought to rescue us from a second-class citizenship vis-à-vis the so-called hard sciences with which we live in close association. Some of our resistance to change and accommodation is also to be attributed to social and economic interest. In the service of such interest, market theory—the subordination of all decision and action to the comprehensive authority of the market—conceals a highly inconvenient reality, which is the promiscuous exercise of power in modern economic life by the large enterprise. How much better if, in our instruction and writing, all economic power be held subordinate to the overriding authority of the market. And some of the failure of economics to respond to changing reality comes from the adverse role of the textbooks. These, in no slight measure, are written not to expose reality but to win acceptance. So, again inevitably, they favor not the changing reality but the previously accepted view. And, alas, there is an inescapable tendency for economists to believe what they teach.

I come to the end of these observations. Self-criticism is not a generally successful professional art form in economics. Accordingly, it is well that I yield to my colleagues on this panel. In this exercise they will not, I assume, share my very considerable, if not wholly unexpected, restraint.

## DISCUSSION

BARRY BLUESTONE, University of Massachusetts-Boston: *The New Industrial State* had the misfortune of being written precisely on the cusp of postwar economic history. In 1967, America was just completing two decades of unparalleled economic expansion; the ratio of imports to GNP suggested that the United States was equally as insulated from the rest of the world's economies as it had been in 1929; and corporate profits were near their post-World War II peak. In this short-lived niche in time, American corporations appeared to be in charge of their own destinies—to be in control of their markets rather than to be controlled by them.

In the following decade, of course, America's preeminence in the world was severely challenged. Imports as a proportion of GNP nearly doubled and profits began a long-term slide that would not end until the beginning of the 1980's. In this environment, the corporation did lose much of the control it once exercised. The important point, however, is that the chief function of the "technostructure" that runs the firm has by no means become obsolete. Along the lines of the central theme in *TNIS*, corporate managers set out to develop new strategies that could effectively supersede, circumvent, or control the market. From the vantage point of 1967, Galbraith would have had to possess extraordinary powers of premonition to predict exactly how managers would eventually cope with the changed environment. But what Galbraith did establish back then—and continues to teach us today—is that in the face of inevitable change, the technostructure will not sit idly by, subject to rigid Lagrangian constraints, but instead will use all of its power to develop fresh strategies to cope with new exigencies.

*TNIS* was an early inquiry into what makes corporate managers tick. Presumably understanding their motivation and their modus operandi would provide a window through which the entire economy could better be appreciated. While this particular perturbation might seem small, its central location in political economy suggested that, peering through it, we might gain more in-

sight not merely about the intricacies of such conventional economic topics as pricing behavior, but also about larger questions: factor shares, the role of government, and ultimately even the prospect for social class collaboration or class conflict. To make this enormous set of issues manageable, critics often reduce the debate over *TNIS* to the more limited question: do corporations profit maximize? Yet, even here, where the traditional economist presumably has the home court advantage, Galbraith's insight survives.

In the environment of record profits during the 1960's, it is altogether reasonable that corporate management could and did allow other goals beside maximum profit to creep into their calculus—maximizing share of market, paying bonuses to itself or to its workforce, or even engaging in hedonistically motivated philanthropic activity (i.e., donating *other* people's money!) But as rates of return on capital slipped from nearly 10 percent in 1965 to an average of less than 6 percent during the entire decade of the 1970's, managers were forced to switch strategy. From long-term profit "satisficing" during a period of extraordinary rates of return, firms had to shift gears to concentrate almost singlemindedly on short-term profit maximization.

While Galbraith admittedly failed to foresee the enormous impact of "the invasive thrust... of Japan and the other Pacific countries" on profits or the degree to which corporate raiding would force the existing technostructures to concentrate on the bottom line, the overall point still stands: when corporations are flush they can afford to pursue other goals beside pure profit maximization and they do; when profits are threatened, they subjugate all other desires and motives to simple short-term greed. In fact, by the end of the 1970's, the shift in strategy from profit satisficing to profit maximizing was sufficiently widespread to account for some of the stagflation that arose. In the face of shrinking demand, corporate managers shifted from a market share maximizing game plan to the standard neoclassical strategy, in the process raising prices.

To cope with declining profits, corporations also set out to abrogate the social contracts that had been negotiated with labor and government beginning in the 1930's—social contracts that had proven functional, according to Galbraith, for maintaining profits during a period of expanding markets and limited international competition. No longer able to boost or even maintain profits through managed pricing policies (and increasingly unable to compete on product quality), managers began to adopt an explicit cost-side strategy to rebuild profit rates.

By almost any measure, they certainly did not succeed in every strategy. OPEC frustrated plans to cut raw material costs, and the Fed eliminated the capital saving gambit. But this left labor cost containment and a radical restructuring of the public sector's relationship to the corporate sector wide open for strategic realignment. The corporate strategies adopted with respect to labor included subcontracting, multinational outsourcing, two-tier wage systems, the use of contingent labor, and the demand for out-right concessions.

A concerted effort on the part of corporate management to cancel an unwritten social contract with government was also launched. It ultimately paid off in terms of lower corporate income taxes and partial deregulation in the areas of labor law, occupational health and safety, and environmental protection.

What is true, of course, is that none of this merely happened. It was consciously planned and executed. Galbraith could not tell us the precise nature of the plans that would be adopted a decade hence, but he did show us where to look. Indeed, given the sea change in the global economic environment, one is actually quite awestruck by just how helpful *TNIS* remains—not as a guidebook to current corporate strategy, but as a framework for understanding the imperatives of the modern corporate planning system.

When *The New Industrial State* first appeared, Robert Solow wrote in review: "The world can be divided into big-thinkers and little-thinkers.... Professor Galbraith makes an eloquent case for big-thinking, and he has

a point. Little-thinking can easily degenerate into mini-thinking or even into hardly thinking at all."

Fortunately, we have Galbraith to thank for keeping our eyes on the prize and not sinking into mini-think or not much think at all.

ROBERT M. SOLOW, MIT: Older members of this audience may remember that I wrote a review of *The New Industrial State* when it appeared. That is no doubt why I am here today. I do not intend, however, to rehash what I said then, except to report that, when I read the review again, it still seemed right to me. That is merely another confirmation of the Law of the Measured Approval of One's Own Recorded Words, of which you see at least two examples on this platform. I wish any proposition in economics were so secure.

There are two things I learned from that experience; I want to mention one of them here as a warning to any fledgling economists who happen to be listening. I can honestly say that I have spent most of my professional life—apart from teaching—either doing or trying but failing to do serious research on serious problems. But every so often the devil makes me write something entertaining, and that is what everyone remembers and wants to Xerox. It illustrates the operation of yet another Law, this time Gresham's. And it exhibits the wisdom of Samuelson's First Rule for Scholars: Never make a joke.

Anyway, there's no point in replaying a twenty-year-old difference of opinion. What I shall do instead is to comment on Galbraith's incremental judgments, on the changes he would now make in his view of the world and of economics in response to the events of the past twenty years. After all, even if the Department of Commerce doesn't know what the GNP *is*, it may be a good guide as to how it is changing.

The first amendment Galbraith suggests is that, while big business has strengthened its control of one aspect of its environment, the federal government, even more than he had initially foreseen, the corporate system has failed to translate its increased power into

a rationally self-interested macroeconomic policy. Even if that is granted, I am not clear if it amounts to a major or a minor amendment of *TNIS*. But I am not inclined to grant the premise in the form in which it is stated. I will certainly agree that there has been a swing toward conservative governments in Britain, Canada, and the United States. In Britain, where bashing the unions is high on the corporate and the party agenda, the outcome has been more to the corporate system's liking. In the United States, the straightforward identification of Reagan conservatism with the agenda of big business seems oversimple, especially when it comes to macroeconomic policy. My impression is that the top executives of large corporations are on the whole traditional fiscal conservatives. The Reagan macroeconomic design—assuming it is a design and not a careless improvisation—is not theirs at all. Big business is Feldsteinian; but Galbraith's colleague *lost* his battle with the conservative political apparatus. It is not the case that the corporate establishment adopted an irrational macro policy; it was not their policy that was adopted.

The other concession Galbraith makes may be much more subversive of the message of *TNIS*. For starters, the success of European and Asian firms in matching (and in many instances surpassing) the level of technology of American counterparts does not reflect well on the "technostructure" who, it will be remembered, are supposed to be motivated strongly by the opportunity to exercise technical virtuosity not to mention market share. One wonders how they get their kicks now. I doubt that this failure can be explained away just by latent bureaucratic tendencies.

There is a more important point, analytically speaking. There are not many industrial firms in the United States of any size and level of technology that can now see themselves as being in control of the market environment. Where import competition is not already a fact of life it is a threat, even in industries—like the manufacture of civilian aircraft—where U.S. firms have been dominant and are still leaders. The availability of several imported substitutes, equal in design and equality, must mean at a mini-

mum that American manufacturers face a more elastic demand and a less controllable one. That does not imply that universal perfect competition is now the rule. If Galbraith had only been urging the brethren to pay more attention to oligopoly and imperfect competition as market forms, he would find it easier to adapt his story to the contemporary world. But he argued for something much stronger—the approximate irrelevance of market forces—and there he has been preempted by events. Big American Business could not keep the internationalization of manufacturing from happening. Whatever feeling of control it used to have must now have been replaced by something more like desperation.

In part, then, Galbraith has just been the victim of bad luck. The rise of the Far Right and the Far East need not have turned out exactly as they did. The Far Right might have remained marginal; there was nothing inevitable about its rise to prominence in the Reagan years. On the other hand, it was probably in the cards that Japan and Taiwan, Korea, Hong Kong, and Singapore would become important players in the world market for sophisticated manufactures. Even so, had private and public attitudes and decisions been slightly different, the extent of Asian penetration of North American markets might have been significantly less. But some of the troubles of *TNIS* were the author's fault. He did have a tendency to see "control" where a less ambitious thinker would have been content to see advantage, and to argue that it resulted from "planning" where a less ambitious person would have been content to see historical accident, the seizing of opportunities offered by current market conditions, and a bit of dumb luck.

Sometimes I think Galbraith was bemused by the more grandiose dreams and claims of some captains of industry. In seeking to oppose them he granted their believability. I did not find that picture credible, and it is even less so now.

I would prefer to adopt a more modest reading of *TNIS*. In that interpretation, Galbraith is making a picturesque and powerful argument against the tendency of some modern economists to treat economic

life as completely autonomous, uninfluenced by political beliefs and politics, independent of social institutions and the norms of behavior they impose on people and cause them to internalize. A profession entranced by the vision of "pure" economics is likely—though not certain—to fall into the habit of modeling the economy abstractly and then forgetting that the result is an abstraction. The next thing you know, you are drawing conclusions from particular models of general competitive equilibrium and applying them to a rather different sort of real world.

You can read *TNIS* as a dramatic monologue in favor of a broader view of economics: one that accepts imperfect competition as a fact of life and views economic motivation and behavior as embedded in a social-political-ideological context. I find that interpretation quite acceptable, much more so than one that takes the book to be a literal description of *The World According to Galbraith*. The main problem with the book in my interpretation is that its author does not seem to understand how hard it is to go down the alternative road with an acceptable degree of rigor. It is not so hard to paint impressionist sketches in that world, a jetty here, a sand-bar there, but you would not care to navigate by one of them. It is a lot harder to do rigorous model building in that more complex world, but I am convinced one has to make a start and learn how. Galbraith should learn to be more discriminating in his abuse.

I want to conclude by just mentioning a problem, often raised by Galbraith, and with justice, that illustrates the difficulties and dangers. I have in mind the social determination of tastes. We—I mean mainstream economists—tend to take tastes as "given." That does not mean we believe them to be genetic or biological or "human nature." It only means that we do not study or question how they originate or how they change. Galbraith insists that it is ridiculous to make and judge policy against a background of given tastes when other actors in the play—firms engaged in advertising and marketing their products, for example—are busily engaged in manipulating tastes as best they can. Sometimes he seems to suggest that

there is no value to "respecting" today's tastes if they are the product of yesterday's adman.

Well, it is not so simple. Today's laws are the product of yesterday's politicians. It would be a pretty drastic step to suggest that therefore there is no virtue in respecting them. You could make a counterargument to that analogy by saying that, in a democracy, today's laws are the residue or trace of the evolving legal principles of the society. But you could make a similar argument about today's tastes: the adman acts on something that is already there. If today's tastes are not to carry much weight in the making and judging of policy, what should take their place? Galbraith caricatures mainstream economists as accepting whatever is as permanent. The caricature strikes back by suspecting that Galbraith proposes to substitute his own preferences for those of ordinary people. I think there is a genuine intellectual problem here; and the mainstream is hiding its discomfort behind glibness. In analogy to the interpretation of law, I suspect policy has to be made and judged by teasing out from changing fashions an interpretation of the bony structure of underlying preferences. How to do that is not easily taught, because it is not easily understood.

F. M. SCHERER, Swarthmore College: In *The New Industrial State*, Galbraith argued that large corporations are complex organizations for which the classical assumption of profit maximization is a gross oversimplification, if not palpably wrong. The problem with which he and many others have struggled is, what to put in the place of profit maximization? On this, we economists have learned a few things over the past two decades, and a return to *TNIS* offers some provocative insights.

In his retrospective, Galbraith observes that the major motivational goals of the large corporation's technostucture include power as well as pecuniary return. Ignored is another 1967 technostucture motivation—the desire of individuals to identify with their organization and to subsume its goals as their own. Such identification, said Galbraith earlier, becomes more important,

the better off workers are materially, the more their employing organization has moved away from control by an identifiable owner group, and the greater the organization's scientific and technical orientation.

On rereading Galbraith's 1967 view of the mature corporation, I was struck that his description fits contemporary Japanese corporations more closely than their American counterparts. In Japan, pecuniary rewards are much less closely correlated with performance than here; individual employees' goals are especially strongly identified with those of the organization; and technocrats play at least as strong a role in shaping organizational goals as they do here. If the Japanese corporation epitomizes Galbraith's paradigm, can we learn something about corporate goal hierarchies by comparing Japanese enterprises against less Galbraithian but "mature" corporations in America? The risk in such a comparison, of course, is that *TNIS* behavioral theory is culture-bound, and that what really differentiates Japanese corporations is not their organization, but more fundamental cultural or economic variables. This is a risk worth taking.

In *TNIS*, mature corporation goals are said to include, in addition to profits, survival, sales growth, and technological virtuosity. From a parallel survey of managers in 1,031 leading Japanese and 1,000 U.S. industrial corporations, Tadao Kagawa and others found "return on investment" to be ranked first as a goal by U.S. enterprises and third by the Japanese. "Higher share prices" were ranked second by the U.S. firms and *eighth* by the Japanese; "market share" ranked third among the U.S. firms and second among the Japanese; while "improving products and introducing new products" occupied fourth place for the U.S. firms and *first* for the Japanese. Plainly, technological virtuosity and sales growth (manifested in market share) are of greater reported importance to Japanese managers, whose organizations conform more closely to the Galbraithian paradigm, than to American managers. The pride of place accorded return on investment and higher stock prices by American managers suggests that, if survey respondents were truthful, they have

not strayed as far from the profit maximization norm as *TNIS* and Galbraith's retrospective imply.

This evidence, however, only pushes the puzzle to a new level. Let us agree with Galbraith of both the 1967 and 1987 versions that uncertainty, the desire for managerial perks, and tendencies toward the accumulation of organizational fat drive a wedge between realized and "maximum" profits. But what, operationally, does profit maximization mean? Surely it means more than equating marginal revenue products with marginal factor prices in the short run of the conventional Economics 1 textbook diagram. Rather, it must mean maximizing the discounted present value of profit streams over time. Once a time element is introduced, it is less clear that the U.S. corporations are the superior maximizers. In the long run, new and better products may do more for profitability than high returns on investment this year. And as modern dynamic limit pricing theory and the theory of learning curve pricing instruct, higher market share today is a strategic variable that can yield higher profits in the future. Whether the technocratic Japanese corporations, with their emphasis on technical virtuosity and market share, are inferior or superior profit maximizers depends at least in part on the sales growth dynamics and discount rates.

It is now recognized that Japanese corporations' capital costs, and hence (presumably) their discount rates, are roughly half those faced by large U.S. corporations. It is at least possible that discount rate differences, rather than differences in organization or culture, explain the Japanese managers' higher regard for product improvement, learning curve pricing, and other market share-enhancing strategies. Whether this, or Galbraithian organizational and cultural variables, explains the observable behavioral differences is a question of the highest research priority.

Japanese and other foreign firms' quest for market share has much to do with the import challenges that have reduced the amount of planning discretion once enjoyed by large American corporations. In modern, technologically virtuosic industries, comparative

advantage does not descend exogenously from a benign Providence; it is fought for. Herein lies a further paradox. Can it be that a long-run strategic view prevails *only* in corporations whose technostucture is insulated to some extent from the gales of competition? If so, the passing from insulated to exposed status may mean *worse*

dynamic performance by American corporations—the opposite of what Galbraith implies. As both business executives and scholars understand these interrelationships better, the insights of *The New Industrial State* may have to be modified, perhaps radically.



# EFFICIENCY WAGES, LABOR RELATIONS, AND FULL EMPLOYMENT<sup>†</sup>

## Relative Wages, Efficiency Wages, and Keynesian Unemployment

By LAWRENCE H. SUMMERS\*

Keynes's *General Theory* in explaining involuntary unemployment advanced the idea that

any individual or group of individuals, who consent to a reduction of money wages relatively to others will suffer a relative reduction in real wages, which is sufficient justification for them to resist it. On the other hand, it would be impracticable to resist every reduction of real wages due to changes in the purchasing power of money, which affects all workers alike. [1936, p. 14]

While modern economic theorists have produced a variety of explanations for the failure of wages to fall in the face of unemployment, Keynes' emphasis on relative wages has not been reflected in most contemporary discussions. This short paper suggests that relative wage theories in which workers' productivity depends primarily on their relative wage provide the best available apparatus for understanding actual unemployment and its fluctuations. Such theories are very closely related to the efficiency wage theories that have received widespread attention in recent years.

Section I motivates and then lays out a simple relative wage model describing the determination of equilibrium unemployment and highlights the fragility of the equilibria

that are likely to result when firms are concerned about their relative wage. Section II develops the close relationship between relative wage models and models that stress the role of insider power in understanding unemployment. Section III shows how efficiency wage models can be extended to account for cyclical unemployment fluctuations once the role of relative wages in influencing worker productivity is recognized.

### I. Relative Wages and Equilibrium Unemployment

For simplicity, consider a labor market in which workers and jobs are homogeneous. In addition to the virtue of tractability, these assumptions remove many of the ambiguities associated with the concept of involuntary unemployment. If the labor market were perfectly competitive and free of information problems, the demand and supply of labor would be equated. In the competitive equilibrium, all firms would pay the prevailing wage, and any worker would be able to immediately obtain work at this wage.

This very simple perfect competition model offers a manifestly inadequate account of the labor markets. Firms do not act as if they face perfectly elastic labor-supply schedules. Small changes in wages do not produce infinite changes in the available supply of labor. In fact, firms focus on variables other than the quantity of labor available to them in setting wages. A large institutional literature has documented that firms go to considerable expense to gain information in order to set an appropriate wage rate relative to other firms in their labor market. In Chicago alone, more than 100 surveys of the wages paid to clerical workers were performed in a single year, while firms went to

<sup>†</sup>*Discussants:* Edward P. Lazear, University of Chicago and Hoover Institution; Juliet Schor, Harvard University.

\*Harvard University, Cambridge, MA 02138, and NBER. The ideas here are developed much more fully in my 1988a paper where a much fuller list of references to prior work is provided. I am indebted to Larry Katz for helpful comments on an earlier draft.

relatively little expense to determine how many clerical workers were unemployed. Most strikingly, even in settings where unemployment is high, firms do not cut wages and they sometimes even raise them.

The natural way for an economist to account for the observation that firms sometimes raise wages even when they are not having trouble staffing their workplace is to postulate that reducing wages in the face of unemployment would reduce profits. Profits may fall when wages are reduced, if reducing wages influences productivity by affecting workers' effort, or by raising the firm's costs of recruiting, training and retaining its labor force. This is the central theme of the burgeoning efficiency literature (surveyed by Joseph Stiglitz, 1986, and Lawrence Katz, 1986) that spells out a variety of mechanisms linking the wages a firm pays to the productivity of its workforce. While the point is rarely emphasized, most efficiency-wage arguments suggest that rather than depending on absolute wages, productivity depends on the relative attractiveness of opportunities inside and outside the firm. Opportunities outside the firm in turn depend on both the wages paid by other firms and the rate of unemployment. Think about stories based on turnover, recruiting, and workers' perceptions of what is fair as examples.

A simple functional form allowing for the possibility that increasing relative wages raises productivity holds that<sup>1</sup>

$$(1) \quad \theta = (w - x)^\alpha \quad 0 \leq \alpha \leq 1,$$

where  $\theta$  measures the effort put forth by the representative worker,  $x$  reflects workers' opportunities in a sense defined precisely below, and  $\alpha$  measures the productivity-enhancing effects of paying higher wages. If  $\alpha = 0$ , efficiency wage considerations are absent. As  $\alpha$  increases, they become more important.

<sup>1</sup>There are a wide variety of devices discussed in the efficiency-wage literature that firms can use to enhance workers' productivity without increasing their wages. In considering the effects of wage changes, I assume that firms have already optimized on all these margins.

The representative firm's problem is to choose a level of wages that minimizes costs per unit of effective labor input,  $w/\theta$ . Differentiating (1) yields the result:

$$(2) \quad w^* = x/(1 - \alpha)$$

which implies that the firm pays workers their opportunity cost if efficiency-wage considerations are absent but generally pays a premium whose magnitude depends on the size of  $\alpha$ .

Characterizing market equilibrium requires a description of how  $x$  is determined. A convenient functional form capturing the idea that outside opportunities depend both on wages paid by other firms and on unemployment is

$$(3) \quad x = w(1 - (1 - b)u),$$

where  $u$  is the unemployment rate,  $w$  is the average wage paid by other firms, and  $b$  reflects their relative importance in determining a worker's outside opportunities. The value of  $b$  in a fully worked out model would depend positively on the utility of leisure, the value of unemployment benefits, and negatively on the duration of unemployment.

Substituting (3) into (2) and requiring that  $w = w$ , since all firms are identical, we obtain a very simple expression for the market equilibrium unemployment rate:

$$(4) \quad u = \alpha/(1 - b)$$

Equation (4) indicates that the equilibrium unemployment rate depends positively on the size of the productivity-enhancing effects of wage increases as reflected in  $\alpha$ , and on the attractiveness of unemployment as reflected in  $b$ . Notice that only in the special and implausible case where  $\alpha = 0$  will there be no unemployment in equilibrium. Notice that the functional form used here has the special and attractive property that the equilibrium level of unemployment does not depend at all on the form of the labor-demand schedule. The labor-demand curve only determines the level of wages. This is an attractive property of the model. It is striking that

real wages have doubled several times over the last century without having a large impact on average unemployment rates.

Substituting plausible parameter values into equation (4), it is clear that only small efficiency-wage effects are needed to account for observed levels of unemployment. Even if  $b = 0$ , a productivity-relative wage elasticity,  $\alpha$ , of only .06 is sufficient to rationalize a 6 percent unemployment rate. For larger values of  $b$ , even smaller efficiency-wage effects are sufficient to rationalize observed levels of unemployment. Furthermore, the image of unemployment suggested by the model also accords with observation in two important respects.

First, the unemployment generated here is involuntary and socially costly. In complex models, it is sometimes difficult to make the concept of involuntary unemployment operational. But here its meaning is clear enough. All jobs and workers are identical. All workers want jobs at the prevailing wages, but only some workers can get them. Furthermore, since workers and firms are identical, the unemployment modeled here does not arise from desirable reallocations of labor power to its highest value use. In this sense, it is consistent with observations highlighting the concentration of unemployment among a small segment of the population that experiences long unemployment durations.

Second, the model is suggestive regarding differences between demographic groups in unemployment rates. Those who value leisure highly and whose turnover is quick are most sensitive to relative wages and will have the highest unemployment rates. Think of teenagers as an obvious example. Alternatively, think of construction workers who can easily move from job to job.

The determination of equilibrium unemployment in a general relative wage model is depicted in Figure 1.<sup>2</sup> The equilibrium unemployment rate has the special property that the representative firm optimizes by

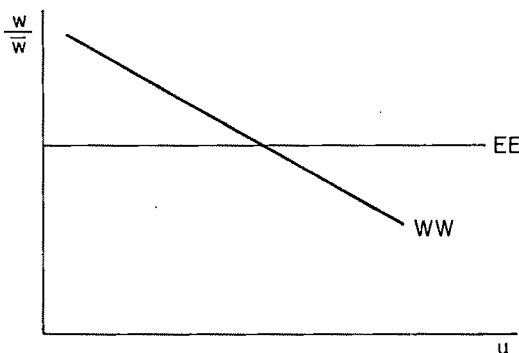


FIGURE 1. RELATIVE WAGES AND EQUILIBRIUM UNEMPLOYMENT

paying the prevailing wage. At lower unemployment rates, the representative firm wants to pay a wage that exceeds that paid by other firms. At higher unemployment rates, it desires to pay a wage that is lower than the wage paid by other firms. Notice that as long as the representative firm would like to pay a higher-than-average wage in the face of completely full employment, the market equilibrium unemployment rate will be positive.

Looking at Figure 1, it is clear that if the two schedules intersect at a narrow angle, small movements in either schedule will have a large effect on equilibrium unemployment. For instance, even if  $\alpha$  is only .03, a relatively modest increase in the value of the "unemployment benefit"  $b$  from .5 to .6 would be sufficient to increase the unemployment rate from 6 to 7.5 percent as the  $WW$  schedule shifted upwards. The sensitivity of the unemployment to small shocks is a consequence of the relative wage model's basic logic. Developments that cause some firms to raise wages have their effects magnified because each firm's optimal wage is a positive function of average wages.

The principle, that concerns with conformity can lead to volatility and instability, is a very general one. It must help to explain why the demand for hula-hoops or Rubik's cube is so much more volatile than the demand for more standard products whose value depends less on whether they are used by others. In the next two sections, I argue

<sup>2</sup>A very similar discussion of the determination of the "natural rate of unemployment" is presented in George Johnson and Richard Layard (1986).

that conformity effects can help to explain why structural and cyclical unemployment vary so widely.

## II. Relative Wages, Insider Power, and Structural Unemployment

The preceding discussion has maintained the assumption that firms are able to set wages in order to maximize their profits. A major theme of recent discussions of unemployment, particularly in the European context, is the idea that wages are set by bargaining, implicit or explicit, between firms and workers. Such bargaining obviously occurs in union contexts. Assar Lindbeck and Dennis Snower's (1987) insider-outsider theories suggest that bargaining may be relevant in nonunion settings as well. Lindbeck and Snower treat insider-outsider theories as an alternative to efficiency-wage theories in explaining unemployment.

From the perspective of the model presented in the first section, it seems more natural to regard them as complementary, mutually reinforcing explanations for unemployment. The relative and efficiency wage considerations stressed in the previous section magnify greatly any effects of bargaining power in two respects. First, in the model developed above, firms reach an interior optimum in setting wages. It is a property of such an optimum that sufficiently small changes in wages have no effect on profits, and larger changes in wages have only second-order effects on profits. This means that in an efficiency wage environment, firms that are forced to pay their workers premium wages suffer only second-order losses. In almost any plausible bargaining framework, this makes it easier for workers to extract concessions.<sup>3</sup>

Second, a key aspect of any relative wage theory is that the optimal wage for a firm to

pay depends positively on the wages paid by other firms. This means that when insiders raise wages at some firms, the effect spills over leading other firms to raise their wages. William Dickens and Katz' (1986) survey of the literature reports some evidence that, contrary to the predictions of at least simple competitive theories, the presence of unions in an industry raises the wages of both union and nonunion workers. Similarly, it is often argued that increases in the minimum wage lead to changes in other wages as well in order to preserve relativities. Relative wage effects on productivity can explain why insider power can create unemployment, even if there are some freely competitive sectors of the economy.

These two points can be illustrated by a simple calculation. Imagine that insiders at a fraction  $\beta$  of all firms have the power to extract a premium of  $\mu$  over wages at unorganized firms. Then the equilibrium unemployment rate may be calculated by solving the equations:

$$(5) \quad w^o = (1 - \mu)w^n$$

$$(6) \quad w^o = (\beta w^n + (1 - \beta)w^0) \times (1 - (1 - b)u) / (1 - \alpha)$$

where  $w^o$  and  $w^n$  represent respectively the wages in the organized and nonorganized sectors. This yields

$$(7) \quad u = (\alpha + \mu\beta) / (1 - b)(1 + \mu\beta).$$

Equilibrium unemployment increases with the size of the organized sector and with the size of the wage premiums it can extract. The results of inserting plausible parameter values are striking. Assume, as before, that  $\alpha = .06$  and  $b = 0$ . Then, if  $\mu$  and  $\beta$  are each equal to .15, insider power will increase the unemployment rate from 6 percent to 8.1 percent. Yet, union firms incur labor costs that are only 6 percent greater than in non-unionized firms because of the productivity-enhancing effects of wage premia.

The role of relative wages explains why unemployment outcomes are so sensitive to small amounts of insider power. This com-

<sup>3</sup>The results of John Abowd (1987) corroborate the efficiency wage hypothesis in this respect. Taking a long horizon into consideration, Abowd finds that surprise increases in wages resulting from collective bargaining agreements reduce firms' market values by much less than the projected increase in labor costs.

ports with the common observation that "corporatist" countries, where labor bargaining is centralized, tend to have lower average rates of unemployment than other nations where bargaining is decentralized.

### III. Cyclical Unemployment

The basic problem in understanding cyclical fluctuations involves isolating the impulses and propagation mechanisms that cause the economy to fluctuate. The relative wage approach to understanding unemployment developed here suggests propagation effects are likely to be strong, and so only small impulses are necessary to account for observed cyclical fluctuations. In particular, the equilibria described in Figure 1 are "fragile"—that is, very sensitive to small disturbances. Small real shocks may have large effects particularly if they are transitory and so do not affect workers' perception of " $x$ " representing outside opportunities.

The relative wage model here suggests that unemployment will be very sensitive to perception errors that might plausibly follow changes in monetary policies. Essentially, misperceptions by workers of average wages shift the *EE* curve in Figure 1 upwards. If relative wage effects are strong, even small misperceptions can have large effects. Imagine that the money stock is reduced, but firms believe that the workers who still have jobs do not yet recognize that equilibrium wages have declined. Then it would not be profitable for them to reduce wages to the level that would be an equilibrium if workers did not misperceive their opportunities. Furthermore, firms that recognized that other firms were not reducing their wages to the new equilibrium level would recognize that they should not either, even if their workers were fully informed.

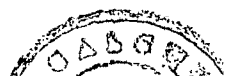
This misperceptions story is reinforced by two further considerations that distinguish it sharply from misperception interpretations of business cycles put forward by classical theorists. First, the central problem here is one of coordination. Notice that it is sufficient to prevent reattainment of equilibrium following a shock, for some firms to suspect that some firms will suspect that some firms

will suspect that...either workers or some firms...will not assume that the new equilibrium is to be attained immediately. The informational requirement for the costless attainment of a new equilibrium is much more than individual rationality—it is common knowledge that all individuals will be rational.

Second, the plausibility of rapid adjustment is further undercut by the observation that, at least in the face of an adverse shock, workers who are perceived as ignorant of the new equilibrium will benefit, in that their wages will not be reduced. This makes it even less likely that transitions between equilibria will occur smoothly. The idea of persistent misperceptions is supported by evidence. John Caskey (1985) demonstrates that inflation was consistently underestimated for ten years during the 1970's and has been consistently overestimated during the 1980's.

The description of wage setting sketched here seems more compelling than the assumption of sticky nominal wages that is contained in "Keynesian" macroeconomics textbooks. Keynesian formulations have been successful in identifying reasons why firms might find it costly or undesirable to vary wages continuously. But most of the reasons they have given for wage rigidity are at least equally plausible as justifications for keeping the level of employment constant and not firing workers during recessions. On the other hand, the misperceptions idea stressed here explains why firms choose to adjust wages slowly and fire workers when adverse shocks come. There is also the further point stressed in some of the implicit contracts literature that layoffs help to educate workers who have jobs about adverse changes in market conditions.

An analogy developed in my earlier paper (1988b) may be helpful in seeing the point of this section. Daylight savings time is purely a change in the "units" used in measuring time. Yet it clearly has real effects in the sense that stores open at a different time relative to the sunrise because of its existence. Why? Probably because most individuals care much more about being on the same time standard as their neighbors than they care about what that time standard is.



Therefore, coordinating actions can succeed in achieving a better outcome in the summertime than the market would generate. Much the same may be true of expansionary policy during recessions.

#### IV. Conclusions

Unemployment, like cancer, is a multifaceted phenomenon that comes in many forms. But one would hope that theory could isolate aspects common to different types of unemployment in different places and times. I suspect that recognizing the role of relative wages in influencing workers' performance will help economists in understanding different types of unemployment. Keynes emphasized the volatility associated with situations where people try to guess the guesses of others in financial markets. This essay has tried to argue that the lesson is a general one applying to labor markets as well.

#### REFERENCES

- Abowd, John M., "Collective Bargaining and the Division of the Value of the Enterprise," NBER Working Paper No. 2137, January 1987.
- Caskey, John, "Modeling the Formation of Price Expectations: A Bayesian Approach," *American Economic Review*, September 1985, 75, 768-77.
- Dickens, William and Katz, Lawrence, "Industry Wage Differences and Industry Characteristics," in K. Lang and J. Leonard, eds., *Unemployment and the Structure of Labor Markets*, London: Basil Blackwell, 1986, 48-89.
- Katz, Lawrence, "Efficiency Wage Theories: A Partial Evaluation" in Stanley Fischer, ed., *NBER Macroeconomics Annual 1986*, Cambridge: MIT Press, 1986, 235-76.
- Keynes, John Maynard, *The General Theory of Employment, Interest and Money*, New York: Harcourt Brace, 1936.
- Johnson, George and Layard, Richard, "The Natural Rate of Unemployment: Explanation and Policy" in O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics*, Vol. 2, Amsterdam: North-Holland, 1986, ch. 16, 921-99.
- Lindbeck, Assar and Snower, Dennis, "Efficiency Wages Versus Insiders and Outsiders," *European Economic Review*, February 1987, 31, 407-16.
- Stiglitz, Joseph, "Theories of Wage Rigidity," in J. L. Butkiewicz et al., eds., *Keynes's Economic Legacy: Contemporary Economic Theories*, New York: Praeger, 1986.
- Summers, Lawrence, (1988a) *Understanding Unemployment*, Cambridge: MIT Press, 1988.
- , (1988b) "Should Keynesian Economics Dispense With the Phillips Curve?," in Rodney Cross, ed., *Hysteresis*, London: Basil Blackwell, forthcoming 1988.

# Unemployment, Labor Relations, and Unit Labor Costs

By JAMES B. REBITZER\*

In his seminal 1943 paper on the political business cycle, Michal Kalecki (1971) argued that "industrial leaders" feared full employment because the economic insecurity created by unemployment was necessary to keep wages low and maintain work intensity and discipline on the shop floor. On the basis of this reasoning, Kalecki concluded that governments would not use demand management policies to achieve permanent "full employment." In terms of current macroeconomic debates, Kalecki had sketched the outlines of a theory of the "neutral" or "natural" rate of unemployment based on the importance of disciplinary unemployment as a regulator of unit labor costs.

Kalecki's pessimism about the prospects for full employment was premised in part on a view of firms in which the threat of dismissal was the central motivational device used by employers, and employees had only a tenuous connection to employers. This assumption may have been appropriate when analyzing labor markets in the United States during the 1930's. Since that time, however, the spread of unions, implicit employment contracts, and large, bureaucratically organized enterprises has resulted in a modern U.S. labor market in which many workers enjoy long job tenure and in which many firms do not appear to rely on dismissal threats as their primary motivational strategy (see David Gordon, Richard Edwards, and Michael Reich, 1982; Sanford Jacoby, 1983; and my forthcoming paper). From this perspective, it is reasonable to ask whether the presence of long-term employment relations alters the regulatory role played by unemployment.

This paper examines the effect that unemployment and long-term employment relations exert on the determination of unit labor costs. The central empirical findings can be

briefly summarized. First, as suggested by Kalecki, movements towards full employment increase the rate of growth of wages and reduce the rate of growth of labor productivity. Second, where long-term employment relations are prevalent, the effect of unemployment on both wage and labor productivity growth is diminished.

## I. Work Effort, Unemployment, and Unit Labor Costs

Much of the spirit of Kalecki's view of unemployment as a regulator of wages and work effort is captured in recently developed efficiency wage models that focus on the firm's problem of eliciting work effort from its employees (see Carl Shapiro and Joseph Stiglitz, 1984; Samuel Bowles, 1985; and Jeremy Bulow and Lawrence Summers, 1986). The premise of these "effort-regulation" models is that the labor exchange is open ended in the sense that specific work activities and work intensity are *not* specified in the employment contract. Rather, these aspects of the terms of employment are determined by the ability of the employer to exercise authority over the employee.

Firms exercise authority by supervising the activities of workers and dismissing those found to be shirking or performing in a substandard manner. The effectiveness of dismissal in eliciting work effort depends upon the cost of dismissal to the worker. The relationship between the cost of dismissal and work effort is often modeled by a work intensity function:

$$(1) \quad L^* = h(W^*), \quad h_{W^*} > 0$$

$$\text{and} \quad h_{W^*W^*} < 0,$$

where  $L^*$  is the effort exerted per labor hour (i.e., work intensity) and  $W^*$  is the cost to the worker of dismissal.  $W^*$  is in turn a function of earnings at the current job ( $W$ ), earnings at alternative jobs ( $\alpha$ ), the probability of experiencing unemployment if dis-

\*University of Texas, Austin, TX 78712. I thank my colleagues Lowell Taylor, Joe Ritter, and Price Fishback for comments on earlier drafts.

missed ( $U$ ) and whatever nonwage income is provided to unemployed workers by the state ( $\beta$ ). Given  $\alpha$ ,  $\beta$ , and  $U$ , the firm chooses a wage that maximizes profits. In most of the published effort-regulation models, the extraction of work effort is considered to be separable from the rest of the production process. In this case, firms set wages so as to minimize the ratio of hourly labor costs to hourly work effort. The equilibrium wage generated by effort-regulation models will generally exceed the market-clearing wage. Equilibrium will therefore be characterized by persistent, involuntary unemployment (Bowles).

The comparative static properties of efficiency wage models suggest that as labor markets tighten, firms' unit labor costs will rise. This property of efficiency wage models follows from the fact that when unemployment falls, firms must offer a higher wage in order to attain any given level of work intensity. Although unit labor costs will tend to rise as the economy approaches full employment, it is not clear if this increase will be caused by rising wages, falling labor productivity, or some combination of the two. For example, in the model developed by Bowles, falling unemployment results in both rising wages and rising productivity. On the other hand, in my (1987) efficiency wage model, falling unemployment may result in rising wages and falling labor productivity.

## II. Long-Term Employment Relations

It is widely accepted among economists that long-term employment relations (LTER) in the United States reflect the presence of labor market structures that create lasting bonds or attachments between workers and their firms. The LTERs are frequently associated with the presence of unions, bureaucratic mechanisms of labor control and implicit contracts. Each of these structures can be expected to reduce the effect that unemployment exerts on unit labor costs.

### A. Unions

The presence of explicit, union contracts can be expected to increase job tenure and

reduce the effect of labor market conditions on wages (Daniel Mitchell, 1980; Richard Freeman and James Medoff, 1984). In addition, there is good reason to believe that in unionized settings, the threat of dismissal is not an important means of eliciting work effort and, therefore, labor productivity is not responsive to external unemployment rates. In unionized firms, supervisors are typically restricted in their ability to dismiss shirking workers, yet a number of empirical studies have found higher levels of productivity in union than nonunion plants (see Charles Brown and Medoff, 1978, and the review in Freeman-Medoff). Freeman and Medoff suggest that these union-productivity effects may result from mechanisms that have little to do with the threat of dismissal. In their view, the union productivity effect is due to the "voice" institutions unions bring to the work place, especially grievance procedures. By offering a means of expressing discontent, these voice institutions improve morale and employer/employee communications and therefore increase the effective labor input per hour of labor employed. The hypothesis that the threat of job loss is not an important motivational device in unionized firms is also supported by Casey Ichniowski's (1986) study of nine unionized paper mills. He found that layoffs do not influence the productivity of workers who remain after the layoffs. To the extent, then, that long-term employment relations are due to the presence of unions, one could expect that they reduce the responsiveness of both wages and labor productivity to changing labor market conditions.

### B. Bureaucratic Control

The "bureaucratic control" approach to LTERs is premised on the observation that large nonunion firms in the postwar period in the United States have established a hierarchical organization in which workers follow career paths determined by the firm's internal job ladders (Richard Edwards, 1979). In these firms there is a well-specified division of labor, and discipline is enforced by impersonal rules governing evaluation, promotion, and dismissal. Under bureaucratic



control, wages and work rules are designed to appeal to notions of procedural "fairness" in order to foster employee commitment to the goals of the firm. Bureaucratic control strategies are closely related to the "partial gift exchange" models developed by George Akerloff (1982) as well as to the various "sociological" aspects of the employment relationship discussed by Robert Solow (1980). Under bureaucratic control, one would expect that wages would be relatively unresponsive to changing labor market conditions. Firms that used a slack labor market to cut wages would be undermining their long-term strategy of cultivating worker loyalty and commitment to the firm. Similarly, bureaucratic motivational schemes rely fundamentally on notions of "fairness" and "legitimacy" rather than dismissal threats to elicit work effort. Therefore one would expect work effort to be unresponsive to changes in the external unemployment rate.

### C. *Implicit Contracts*

In the context of efficiency wage models, it can be demonstrated that implicit contracts reduce the sensitivity of unit labor costs to changing labor market conditions. It has long been recognized that high turnover costs resulting from investments in firm-specific training can move firms to offer implicit contracts guaranteeing employment at a wage that rises with job tenure (Arthur Okun, 1981). Implicit contracts can also arise where workers are more risk averse than firms (see Martin Neil Bailey, 1974). In this situation, workers will prefer a less volatile income stream to a riskier income stream. Firms will gladly offer to provide insurance against income fluctuations in exchange for a lower wage rate.

Workers who invest in firm-specific training (or who are more risk averse) suffer a greater loss upon dismissal than other workers paid an equivalent wage. Edward Lazear (1981) has demonstrated that even in the absence of firm-specific training or risk aversion, implicit contracts guaranteeing employment over time at an increasing wage will increase the cost of job loss to the worker. Lazear and Robert Moore's (1984)

study suggests that the slope of observed age-earnings profiles is significantly affected by these sorts of work incentives. Thus one would expect, on the basis of implicit contract theory, that long-term employment ought to be associated with high costs of job loss.

Using a simple efficiency wage model based on equation (1), it can be demonstrated that an increase in the cost of job loss to the worker,  $W^*$ , reduces the responsiveness of unit labor costs to changing labor market conditions.<sup>1</sup>

### III. Estimating the Effect of Unemployment on the Growth of Wages, Labor Productivity, and Unit Labor Costs

The comparative static properties of effort-regulation models suggest that as the economy moves toward full employment, unit labor costs will rise. Diverse theories of long-term employment relations on the other hand suggest that the presence of long-term employment relations ought to reduce the effect falling unemployment exerts on unit labor costs. This section presents the results of empirical explorations of these hypotheses using annual data from 2-digit U.S. manufacturing industries over the period 1961-80.

In order to examine the effect of falling unemployment on unit labor costs, it is necessary to examine the effect unemployment exerts on both wages and labor productivity. Following Mitchell, I estimated a wage equation that regressed the annual rate of growth of nominal wages ( $CHEARN$ ) against the inverse of the civilian unemployment rate ( $UR1$ ) and the lagged rate of growth of the Consumer Price Index ( $INFL_{t-1}$ ). In addition, I included as explanatory variables, the rate of unionization in the industry ( $UNMEM$ ) and variables measuring the ability to pay higher wages (i.e., the current

<sup>1</sup>This conclusion follows from the diminishing marginal effectiveness of costly dismissal in eliciting work effort. The effect that increasing  $W^*$  has on the unemployment/unit labor cost relationship is discussed more fully in an appendix to this paper which is available from the author upon request.

and lagged rate of growth of labor productivity,  $CPROD1_t$  and  $CPROD1_{t-1}$ , respectively). Variations in the effect of unemployment on wage and labor productivity growth were captured by interacting the unemployment variable ( $UR1$ ) with a measure of the prevalence of long-term employment relations in the industry ( $JT^*$ ). Similarly the lagged inflation term in the wage equation ( $INFL_{t-1}$ ) was interacted with the measure of long-term employment relations in the industry.<sup>2</sup>

The measure  $JT^*$  was derived from estimates of average industry job tenure. However, microeconomic studies of the determinants of job tenure indicate that it is strongly influenced by factors such as age and gender that are not directly related to the labor market structures described above (see my 1986 paper). Consequently, this study made use of a standardized measure of industry job tenure that controls for cross-industry variation in personal characteristics.<sup>3</sup>

Data limitations prohibited the estimation of a separate wage growth equation for each industry. Therefore the following pooled time series and cross-sectional equation was estimated:

$$(2) \quad CHEARN_{it} = a_0 + a_1(UR1)_t + a_2(UR1)_t(JT^*)_t + a_3(INFL)_{t-1} + a_4(INFL)_{t-1}(JT^*)_t + a_5(UNMEM)_t + a_6(CPROD1)_{it} + a_7(CPROD1)_{it-1} + v_{it},$$

where  $v_{it}$  is the error term for industry  $i$  at time  $t$ .

In order to examine the effect of unemployment on work effort, I estimated a pro-

ductivity growth equation that used as its dependent variable the annual rate of growth of output per labor hour in the industry ( $CPROD1$ ). The explanatory variables included the inverse of the civilian unemployment rate ( $UR1$ ), the rate of growth of the ratio of capital services to labor hours ( $CKL$ ), the rate of change of the rate of capacity utilization in the industry ( $CWCU$ ), and the rate of unionization in the industry ( $UNMEM$ ). As additional measures of the quality of the capital stock, the equation also included a variable measuring the average age of the real gross capital stock ( $AGE-STOCK$ ) and a variable measuring the fraction of the gross capital stock that is new investment ( $INVRATIO$ ). Two variables were also introduced into the productivity equation to capture secular trends in the rate of growth of labor productivity. These were  $TDUM73$ , a dummy variable equal to one for the years 1973–80 and  $TDUM66$ , a dummy variable equal to one for the years 1966–72. Consistent with the efficiency wage approach, I also included as explanatory variables the rate of growth of real wages in the industry ( $CRHEARN$ ) and the growth of real wages interacted with the measure of LTERs ( $(CRHEARN)(JT^*)$ ). The resulting productivity growth equation can be written:

$$(3) \quad CPROD1 = \alpha_0 + \alpha_1(UR1)_t + \alpha_2(UR1)_t(JT^*)_t + \alpha_3(CKL)_{it} + \alpha_4(CRHEARN)_{it} + \alpha_5(CRHEARN)_{it}(JT^*)_t + \alpha_6(CWCU)_{it} + \alpha_7(AGESTOCK)_{it} + \alpha_8(INVRATIO)_{it} + \alpha_9(UNMEM)_t + \alpha_{10}(TDUM73)_t + \alpha_{11}(TDUM66)_t + \epsilon_{it},$$

where  $\epsilon_{it}$  is the error term for industry  $i$  at time  $t$ .

Equations (2) and (3) constitute a system of equations and were estimated using conventional two- and three-stage least squares techniques. The estimates of  $a_1$ ,  $a_2$ ,  $\alpha_1$ , and  $\alpha_2$  in equations (2) and (3) indicate that an

<sup>2</sup>Many of the same aspects of long-term employment relations which make wages relatively unresponsive to changing labor market conditions will also make wages more responsive to changing rates of inflation. See my forthcoming paper for a discussion of this issue.

<sup>3</sup>Details on the construction of  $JT^*$  are presented in the appendix (see fn. 1).

TABLE 1—PREDICTED EFFECT OF AN INCREASE  
IN THE CIVILIAN UNEMPLOYMENT RATE  
FROM 4.8 TO 6.2 PERCENT

Change in Annual Rate of Growth of	LTER Industries		
	Low	Mean	High
Nominal Wages	-0.588	-0.373	-0.184
Labor Productivity	0.328	0.262	0.140
Nominal Unit Labor Costs	-0.916	-0.635	-0.324

Note: Details on calculations provided in the appendix available upon request.

increase in unemployment has the effect of slowing the rate of growth of wages while enhancing the rate of growth of labor productivity.<sup>4</sup> Moreover, the effect of unemployment is diminished the greater the prevalence of LTERs in the industry. The magnitude of these unemployment effects are illustrated in Table 1.

Table 1 presents estimates of the cumulative effect that a one-time permanent increase in the civilian unemployment rate from 4.8 percent (average unemployment rate, 1960–69) to 6.2 percent (average, 1970–79) has on the growth of labor productivity and wages in low, mean, and high LTER industries.<sup>5</sup> In low LTER industries, this movement in the unemployment rate will, all else equal, reduce the rate of growth of nominal wages by 0.588 percentage points and add 0.328 percentage points to the annual rate of growth of productivity. The net effect is a reduction in nominal unit labor cost growth rates of 0.916 percentage points. In high LTER industries, the change in

unemployment will reduce nominal wage growth by only 0.184 percentage points and increase productivity growth by only 0.140 percentage points. The net effect is therefore a 0.324 percentage point reduction in nominal unit labor cost growth rates.

It is clear from the table that, all else equal, LTERs become relatively expensive during periods of high unemployment and relatively inexpensive during periods of low unemployment. Using the total multipliers calculated from the wage and labor productivity equations, it is instructive to calculate the unemployment rate at which nominal unit labor costs begin to grow more rapidly in long- than short-tenure industries. Assuming that the inflation rate was 7.1 percent per year (the average for the 1970–79 period) and that long- and short-tenure industries were identical in all other respects, then unit labor costs would grow more rapidly in long- than short-tenure industries when the unemployment rate exceeded 5.7 percent. However, if the inflation rate were 10 percent, then unit labor cost growth rates would be higher in long-tenure than short-tenure industries when the unemployment rate exceeded 4.0 percent.<sup>6</sup>

#### IV. Conclusion

The results presented above have implications for the study of the microfoundations of macroeconomics. The findings suggest that some of the barriers towards achieving full employment identified by Kalecki *do* appear to operate in the modern U.S. economy. As the economy approaches full employment, the tightening of labor markets causes wage growth to increase and labor productivity growth to slow. However, these adverse effects of low unemployment are considerably reduced where firms have established long-term employment relations with their employees. A provocative implication of these findings is that long-term employment

<sup>4</sup>Three-stage least squares estimates (and *t*-statistics) for  $a_1$ ,  $a_2$ ,  $\alpha_1$ , and  $\alpha_2$  were 21.525 (3.456), -1.160 (-2.544), -24.601 (-2.780), and 1.361 (2.478), respectively. The mean value of  $UR1$  was 0.190 and the mean value of  $JT^*$  was 12.025. Further details concerning data, estimation techniques and results are presented in the appendix.

<sup>5</sup>Low, mean, and high tenure industries were given  $JT^*$  values of 8.452, 12.025, and 15.316, respectively. The derivation of the total multipliers used to construct Table 1 are presented in the appendix.

<sup>6</sup>The relationship between unit labor cost growth rates, inflation rates, and unemployment rates in high- and low-tenure industries is derived in the appendix.

relations of the sort seen in the modern U.S. economy may provide an appropriate microeconomic foundation for a low unemployment macroeconomic regime. Put differently, increased use of long-term employment relations may reduce the "neutral" or "natural" unemployment rate.

The results reported here also have implications for what might be called the macro foundations of microeconomics. It seems to be the case that long-term employment relations become costly relative to short-term employment relations when unemployment and inflation rates are high. Prolonged periods of high unemployment (particularly when accompanied by high rates of inflation) can therefore be expected to erode an otherwise stable system of long-term employment relations. The effect of changing macroeconomic conditions on the stability of long-term employment relations in the United States and elsewhere is a promising area for future research.

## REFERENCES

- Akerlof, George A., "Labor Contracts as a Partial Gift Exchange," *Quarterly Journal of Economics*, 1982, 97, 543-69.
- Baily, Martin Neil, "Wages and Employment Under Uncertain Demand," *Review of Economic Studies*, January 1974, 41, 1043-63.
- Bowles, Samuel, "The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian, and Marxian Models," *American Economic Review*, March 1985, 75, 16-36.
- Brown, Charles and Medoff, James, "Trade Unions in the Production Process," *Journal of Political Economy*, June 1978, 86, 355-78.
- Bulow, Jeremy I. and Summers, Lawrence H., "A Theory of Dual Labor Markets with Applications to Industrial Policy, Discrimination, and Keynesian Unemployment," *Journal of Labor Economics*, July 1986, 4, 376-414.
- Edwards, Richard, *Contested Terrain: The Transformation of Work in the Twentieth Century*, New York: Basic Books, 1979.
- Freeman, Richard B. and Medoff, James L., *What Do Unions Do?*, New York: Basic Books, 1984.
- Gordon, David, Edwards, Richard and Reich, Michael, *Segmented Work. Divided Workers*. London: Cambridge University Press, 1982.
- Ichniowski, Casey, "The Economic Performance of Survivors after Layoffs: A Plant Level Study," NBER Working Paper No. 1807, January 1986.
- Jacoby, Sanford M., "Industrial Labor Mobility in Historical Perspective," *Industrial Relations*, Spring 1983, 20, 261-82.
- Kalecki, Michal, "Political Aspects of Full Employment," in *Selected Essays in the Dynamics of the Capitalist Economy: 1933-1970*, Cambridge: Cambridge University Press, 1971.
- Lazear, Edward P., "Agency, Earnings Profiles, Productivity, and Hour Restrictions," *American Economic Review*, September 1981, 71, 606-20.
- \_\_\_\_\_, and Moore, Robert, "Incentives, Productivity and Labor Contracts," *Quarterly Journal of Economics*, May 1984, 99, 275-96.
- Mitchell, Daniel J. B., *Unions, Wages, and Inflation*, Washington: The Brookings Institution, 1980.
- Okun, Arthur, *Prices and Quantities: A Macroeconomic Analysis*, Washington: The Brookings Institution, 1981.
- Rebitzer, James B., "Establishment Size and Job Tenure," *Industrial Relations*, Fall 1986, 25, 292-302.
- \_\_\_\_\_, "Unemployment, Long-Term Employment Relations and Productivity Growth," *Review of Economics and Statistics*, November 1987, 69, 627-35.
- \_\_\_\_\_, "Efficiency Wages and Implicit Contracts: An Institutional Evaluation," in Robert Drago and Richard Perlman, eds., *Microeconomic Issues in Labor Economics: New York Approaches*, Sussex: Wheatsheaf Books Ltd., forthcoming.
- Shapiro, Carl and Stiglitz, Joseph E., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, 74, 433-44.
- Solow, Robert, "On Theories of Unemployment," *American Economic Review*, March 1980, 70, 1-11.

# Labor Discipline and Aggregate Demand: A Macroeconomic Model

By SAMUEL BOWLES AND ROBERT BOYER\*

The neoclassical theory of employment and output may be characterized by its two most basic abstractions: the acceptance of Say's law and the representation of labor as a commodity like any other input. In practice, Say's law is nothing more than the assertion that product market clearing will be achieved through some combination of price level and interest rate effects, and that the process by which these effects work is sufficiently rapid and regular to justify abstracting from other possible consequences of the failure of markets to clear, such as quantity adjustments. The representation of labor as a commodity denies its more obvious status as human activity motivated in part by the intentions of the worker, and disciplined, if possible, by the employer; in practice it entails the assertion that the amount paid for an hour's work, and the amount of work done in an hour are taken by the firm as exogenously determined.

In this essay (and a series of related papers), we integrate the two absences in the neoclassical theory: the Keynes-inspired analysis of aggregate demand and the Marx-inspired analysis of the problem of wage setting labor discipline. Our model expresses a very simple underlying logic: the distribution of income is a key determinant of aggregate demand through its effects on savings and investment; the distribution of income is in turn the outcome of a class conflict over work and pay in which the balance of class forces is dependent on the level of employment and hence on the level of aggregate demand. Through its effect on the bargaining power of workers and em-

ployers, the level of government redistributive expenditure will influence both the distribution of private incomes and, independently of this, the level of aggregate demand; it will be modeled explicitly and generated endogenously. As we will see, taking account of the effect of the wage on both aggregate demand and the endogenous determination of output per labor hour, the level of employment may respond either positively or negatively to changes in the wage rate, giving rise to what we term a wage-led or profit-led employment regime, respectively.

## I. Equilibrium Wages and Employment Rents in a Labor Extraction Model

Production of a single commodity using homogeneous labor is described jointly by a production function and labor extraction function, the latter representing a dismissal threat system of labor relations governing the pace of work:

$$(1) \quad Q = qhe \quad e = e(w_c)$$

$$\text{and} \quad w_c = w - (hw_a + (1-h)w_u),$$

where  $Q$  is total output,  $q$  is the level of output per unit of work performed,  $h$  is the fraction of a given level of labor supply (normalized as unity) that is employed,  $e$  is work effort performed per hour of labor employed,  $w_c$  is the worker's expected income loss associated with losing his or her job (the employment rent or cost of job loss),  $w_a$  is the worker's expected income in alternative employment and  $w_u$  is the level of income-replacing social benefits that the worker may expect to receive should the job be terminated. We assume throughout that  $w > w_u$ . The labor extraction function,  $e(w_c)$ , is derived from the workers' optimal choice of work effort given the cost of job loss; under quite general conditions it will be the

\*Department of Economics, University of Massachusetts, Amherst, MA 01002, and Centre d'Etudes Prospectives d'Economie Mathématique Appliquées à la Planification, 142, rue Chevaleret, Paris. This essay reports on research conducted as part of the World Institute of Development Economics Research (Helsinki) project on global macroeconomic modeling.

case that wage increases induce greater effort, though at a diminishing rate:  $e' > 0$  and  $e'' < 0$ . The average product of work,  $q$ , is assumed to be independent of the level of employment, the result of unutilized capital stock of homogeneous quality and constant returns to scale; for convenience  $q$  is set equal to unity and dropped from further consideration.

As the microeconomics of this and analogous systems have been explored in a number of our related papers (for example, 1987), and elsewhere by George Akerlof and Janet Yellen (1986), Joseph Stiglitz (1987), and others, it will be sufficient to comment that the competitive profit-maximizing employer will vary the wage offered to labor in order to minimize  $w/e$ , the cost of an effective unit of labor, balancing the direct cost of the wage payment against its positive effect on the cost of job loss and hence on effort. Assuming that the firm regards all of the components of  $w_c$  except  $w$  as exogenous, the first-order conditions entailed by this optimization process are that the firm's optimal (cost-minimizing) wage,  $w^0$ , is such that the marginal effectiveness of a wage change on effort must be equal to the average effort per dollar of wage payment, or

$$(2) \quad e' = e/w^0 \quad \text{or} \quad w^0 = e(w_c)/e'(w_c) \\ = w^0(w_u, w_a, h).$$

The firm's optimal wage  $w^0$  rises with the employment rate: as unemployment falls the best the employer can do is to pay workers more, offsetting the workers' greater employment security by increasing the hourly wage loss associated with job termination. Differentiation of (2) indicates that

$$(3) \quad dw^0/dh = (1 - e'/we'')(w_a - w_u).$$

The firm's first-order conditions (2) plus the labor market equilibrium condition that homogeneous labor must receive the same wage (or  $w = w_a$ ) allow us to define the equilibrium wage,  $w^*$ . Taking account of both conditions, the change in the equilibrium

wage induced by a change in employment is

$$(4) \quad dw^*/dh = \frac{dw^0/dh}{1 - h(1 - e'/we'')}.$$

The inverse of the denominator may be interpreted as a multiplier reflecting the fact that each employer's wage increase raises the worker's alternative wage, lowering the cost of job loss and hence inducing further wage increases. Equation (4) describes a high-employment wage explosion: as  $h$  approaches a limit employment level,  $h_{\max}$ , what we term the wage explosion multiplier becomes infinite, driving equilibrium wages to infinity. From (4) it can be seen that

$$(5) \quad h_{\max} = (1 - e'/we'')^{-1}.$$

This wage explosion employment limit is clearly less than unity, as  $e'$  and  $w$  are positive and  $e''$  is negative. Thus we may summarize the production and labor extraction side of the model by the wage determination function:

$$(6) \quad w^* = w^*(h)$$

with  $dw^*/dh > 0$  for  $w > w_u$  and  $h < h_{\max}$

which may be interpreted as a condition for the stationarity of  $w$ . Not surprisingly, along the equilibrium wage function, profit per hour of labor employed,  $e - w$ , falls as the employment rate rises.

It will be important to note for what follows that if we confine ourselves to labor market equilibria, such that  $w = w_a$ , the cost of job loss is

$$(7) \quad w_c = (1 - h)(w - w_u).$$

Thus, considering a given wage rate, two effects will determine the impact of employment rate rises on the movement of the total profits,  $r$  (which given a particular level of capital stock normalized at unity, is also the profit rate): increasing  $h$  will tend to raise  $r$  if profit per hour of labor is positive, but will tend to lower  $r$  through the negative effect

f increased labor market tightness on the cost of job loss  $-(w - w_u)$  and hence on labor intensity. Thus

$$8) \quad r = h(e - w)$$

$$\text{and} \quad dr/dh = e - w - he'(w - w_u),$$

indicating that for positive profits as  $h$  rises  $w$  will first rise and then fall, describing a high-employment profit squeeze.

Correspondingly, the effect on total profits of a wage increase (for a constant  $h$ ) reflects a positive labor intensity effect operating indirectly via the impact of the wage increase on cost of job loss  $(1 - h)$ , offset by a negative direct wage effect. It will be unambiguously negative in the neighborhood of the firm's optimum:

$$9) \quad dr/dw = h \{ e'(1 - h) - 1 \} < 0$$

for  $w > w_{r, \max}$

The bracketed expression on the right will be zero at the wage,  $w_{r, \max}$ , which maximizes total profits for a given level of employment, and hence would be set by a single profit-maximizing cartel; from (1) and (2) it can be shown that the atomistically competitive wage  $w^*$  exceeds  $w_{r, \max}$  for reasonable values of  $h$ .)

## II. Wage-Led and Profit-Led Employment in an Aggregate Demand Model

Turning now to the demand side of the model, we seek a stationarity condition for  $h$  based on the demand for labor. We will simplify matters greatly and highlight the effects of income distribution on aggregate demand by making both savings and investment depend on the level of profits. Investment will thus depend both on the activity level of the economy ( $h$ ) and the cost conditions facing employers ( $e - w$ ). We assume that the fraction of profits saved is  $s$ , and that all wages are consumed. We abstract from taxation, assuming that the government borrows an amount equal to autonomous borrowing,  $b_0$ , plus the endogenously generated level of income-replacing social pay-

ments  $((1 - h)w_u)$ . Thus (abstracting from net exports) the components of aggregate demand and their relationship to aggregate supply may be expressed:

$$(10) \quad i = i_0 + i_r(h(e - w))$$

$$c = h(w + (1 - s_r)(e - w));$$

$$b = b_0 + (1 - h)w_u \quad \text{and} \quad he = i + c + b$$

where  $i$ ,  $c$ , and  $b$  are, respectively, investment, privately financed consumption, and government spending (borrowing), all expressed as a fraction of the capital stock.

We may express the demand for labor as the total demand for goods divided by the (endogenously generated) average product of labor, or equivalently as the level of employment at which intended investment is equal to intended private savings ( $s$ ) minus government borrowing. Both imply that firms are demand constrained: they will hire more labor if excess demand for goods ( $D_x$ ) is positive. More generally, the rate of change of employment,  $\dot{h}$  is

$$(11) \quad \dot{h} = H(D_x)$$

$$\text{where} \quad D_x = i + b - s,$$

$$H' > 0 \quad \text{and} \quad H(0) = 0.$$

Using (10) and (11) we can write excess demand as an implicit function in  $h$  and  $w$ , and interpret this function as a stationarity condition for  $h$  when  $D_x(h, w) = 0$ .

This product market-clearing condition indicates the effect of a wage change on the equilibrium level of employment: a wage-led employment regime is said to obtain if this effect is positive. The occurrence of a wage-led or (its converse) a profit-led employment regime will depend on the relative importance of the high-employment profit squeeze, the size of the unemployment benefit, and the responsiveness of both savings and investment to the profit rate. For example, one variant of a wage-led regime will obtain if the effect of a wage increase is to increase consumption more than it lowers investment (thus generating excess demand) and if an

employment increase will have the effect of reducing excess demand (through its effect on profits and hence on savings and investment as well as on government borrowing). In this case, in order for product markets to clear, the excess demand induced by a wage increase must be offset by an excess demand-reducing employment increase; given the demand constrained nature of the firms' employment decision, their autonomous actions will in this case generate the equilibrating movement.

### III. Wage-Led and Profit-Led Employment Regimes

The slope of the product market-clearing condition,  $dh^*/dw$ , is evidently worth closer inspection, for it determines whether an employment regime will be wage-led or profit-led. From (8) through (12) we may write

$$(12) \quad dh^*/dw = (dDx/dw)/(-dDx/dh) \\ = \frac{dr/dw(i_r - s_r)}{-dr/dh(i_r - s_r) + w_u}$$

Because (from (9))  $dr/dw$  is negative near the firm equilibrium, the sign of the numerator depends solely on whether investment responds more or less than savings to a change in profits: we term the former a case of investment-led aggregate demand; the latter is consumption-led. The sign of the denominator depends not only on  $(i_r - s_r)$ , but also on the size of the unemployment benefit, and on the effect of variations in employment on profits and hence on the labor extraction function. From (8) we know that  $dr/dh$  may be of either sign and because of the high-employment profit squeeze is likely to be negative for high levels of  $h$ . On the basis of the savings investment behavior and the importance of the profit squeeze relative to the size of the unemployment benefit, we can distinguish the four cases presented in Figure 1.

The wage-led employment regime based on consumption-led aggregate demand described above appears in the upper right cell and is illustrated in Figure 2. As saving is

Production and labor relations:	Aggregate Demand is	
	Investment-led ( $i_r > s_r$ )	Consumption-led ( $i_r < s_r$ )
Insignificant profit squeeze	WAGE-LED i.e., $dh^*/dw > 0$ because $dDx/dw < 0$ and $dDx/dh > 0$	WAGE-LED i.e., $dh^*/dw > 0$ because $dDx/dw > 0$ and $dDx/dh < 0$
Significant profit squeeze	PROFIT-LED i.e., $dh^*/dw < 0$ because $dDx/dw < 0$ and $dDx/dh < 0$	PROFIT-LED $dh^*/dw < 0$ because $dDx/dw > 0$ and $dDx/dh > 0$

FIGURE 1. WAGE-LED AND PROFIT-LED EMPLOYMENT REGIMES

more profit responsive than is investment  $dDx/dw$  is positive; and as the full-employment profit squeeze is insignificant,  $dDx/dh$  is negative yielding a positively sloped market-clearing function. (From (9) and (12) it can be seen that  $h^*(w)$  is vertical at the cartel profit-maximizing wage,  $w_{max}$ .) In Figure 2, horizontal arrows indicate the sign of excess demand (and hence, by (11), the sign of  $dh$ ); the vertical arrows indicate the out of equilibrium adjustment of the wage. The values  $w^{**}$  and  $h^{**}$  represent a stable unemployment equilibrium in which  $dh^*/dw > 0$ , that is, a wage-led employment regime.

An increase in the employment rate will as we have seen, change the sign of  $dr/dh$  from positive to negative; this high-employment

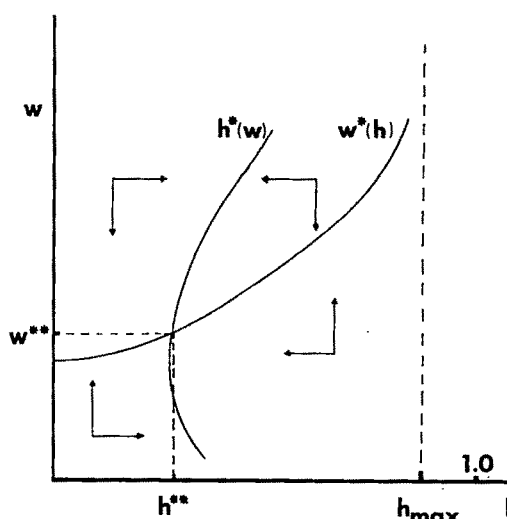


FIGURE 2. A WAGE-LED EMPLOYMENT REGIME



ment profit squeeze may eventually dominate  $dDx/dh$ , changing the sign of  $dh^*/dw$ . The critical value  $h_{lim}$ , at which  $dDx/dh$  is zero and hence the sign of  $dh^*/dw$  changes, is the limit of the wage-led employment regime. (The dividing line between the "insignificant" and "significant" profit squeeze in Figure 1 is, of course,  $h_{lim}$ .) Equilibria for  $h > h_{lim}$  are necessarily unstable as long as savings is more responsive to profits than is investment.

Interestingly, in the consumption-led aggregate demand case, this limit will occur at a higher level of employment the larger is the unemployment benefit. This is because with a large unemployment benefit, employment increases will tend to enhance net savings ( $s - b$ ) by inducing a larger reduction in government borrowing to offset the employment-increase-induced rise in privately financed consumption, and because the high-employment labor intensity squeeze will be attenuated. It is clear from (8) that as  $w_u$  approaches  $w$ , the high-employment profit squeeze will vanish, as it is based on the increase in worker security as employment rises. Thus there need not be a regime shift from wage-led to profit-led over the feasible range of employment levels.

Consider, now, the investment-led case, in which  $i_r > s_r$ . Here  $dDx/dw$  is unambiguously negative, as the negative wage effects on investment dominate the positive wage effects on consumption. In this case, a stable equilibrium will exist when  $dDx/dh$  is negative, which for  $i_r > s_r$  will occur for sufficiently high levels of employment and unemployment insurance, illustrated in Figure 3. We may call this employment regime profit-led, as the equilibrium level of employment is a negative function of the wage rate.

The upper unemployment limit of the profit-led employment regime is set by the wage explosion dynamic outlined above:  $h_{rzero}$  may be defined as the level of  $h$  for which  $(e - w)$  goes to zero. Or, to take an extreme but simple alternate case, we may assume that investment responsiveness to profits is infinite at some given level of profit,  $r$ , (perhaps due to a global hypermobility of capital that establishes a common world profit rate). In this case, we may simply

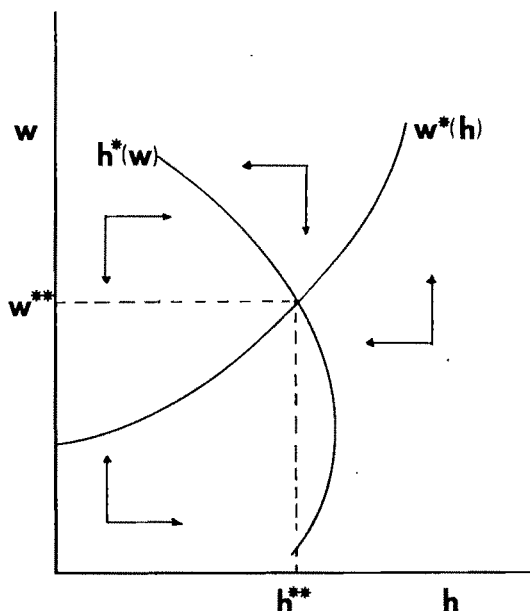


FIGURE 3. A PROFIT-LED EMPLOYMENT REGIME

ignore all the non- $i_r$ -related terms in (12), and note that the resulting expression for  $dh^*/dw$  is an isoprofit function written in  $h, w$  space defining the level of profits  $r$ . Given the assumptions, product market clearing can only be achieved at one level of profit, the level of employment being determined as that which generates this level.

Our last case—wage-led employment in an investment-led demand regime—is at once paradoxical and possibly of some concrete relevance. In this configuration, illustrated in Figure 4, an increase in wages and employment generate, respectively, a decrease and an increase in aggregate demand. The resulting positively sloped market-clearing function may have more than one intersection with the equilibrium wage function, as is shown. Point *a* describes a wage-led employment regime, but with this element of paradox: to generate a simultaneous increase in the equilibrium wage and employment level, a *downward* shift in the wage function is required. We may describe this as a collective wage-led employment regime to high-

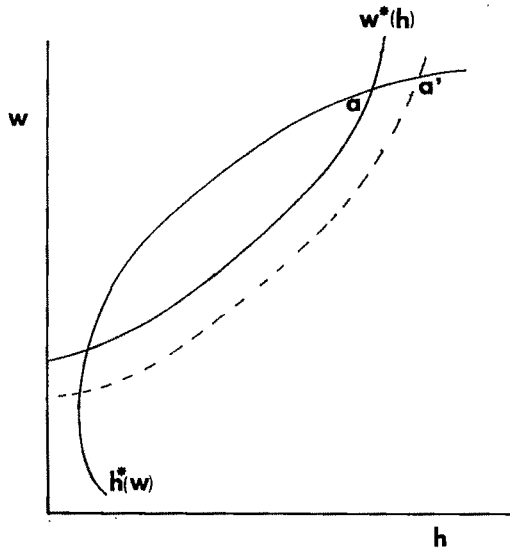


FIGURE 4. ANOTHER WAGE-LED EMPLOYMENT REGIME

light the fact that the joint expansion of employment and wages is obtained through organized wage restraint in the form of workers agreeing not to use the increase in bargaining power which increased employment levels confer on them.

#### IV. Labor Relations and Macroeconomic Theory

What has been the upshot of our simultaneous rejection of both Say's law and the representation of labor as a commodity?

First, our model exhibits a high-employment profit squeeze, giving rise to a limit level of employment which is less than full employment. Beyond this employment level, aggregate demand expansion policies will be ineffective in the absence of institutional changes in the organization of the labor market and work process designed to replace the dismissal threat based system of labor control.

Second, we identify quite general conditions supporting a wage-led employment regime. However, even if we were to adopt assumptions which would appear quite

favorable to a wage-led employment regime—exogenous investment demand and no savings out of wage income—a wage-led employment regime is not guaranteed, and because of the high employment profit squeeze will generally not obtain at high levels of employment.

Third, reductions in the difference between the wage and income-replacing social payments such as unemployment insurance will expand the range of employment rates over which wage-led employment regimes may obtain. Thus there may be a symbiotic quality to the social democratic program that has insisted that wage increases in conjunction with the expansion of the welfare state may enhance employment.

Fourth, even if investment is highly responsive to profits, a wage-led employment regime may nonetheless obtain, but enjoying of its obvious advantages—a simultaneous increase in equilibrium wages and employment—may require collective action by unions, employers, or states to insure wage restraint.

Fifth, at high levels of employment and with investment highly responsive to profitability, wage-led growth is precluded, as aggregate demand comes to be dominated by the negative effect of wage increases on investment.

#### REFERENCES

- Akerlof, George and Yellen, Janet, *Efficiency Wage Models of the Labor Market*, Orlando: Academic Press, 1986.
- Bowles, Samuel and Boyer, Robert, "A Wage-Led Employment Regime: Income Distribution, Labor Discipline, and Aggregate Demand in Welfare Capitalism," in S. A. Marglin, ed., *The Golden Age of Capitalism*, Oxford: Oxford University Press 1989.
- Bowles, Samuel, "The Production Process in a Competitive Economy: Walrasian Marxian, and Neo-Hobbesian Models," *American Economic Review*, March 1985 75, 16–36.
- Stiglitz, Joseph, "The Causes and Consequences of the Dependence of Quality on Price," *Journal of Economic Literature* March 1987, 25, 1–48.

# WHAT DO WE KNOW ABOUT CONSUMPTION AND SAVING, AND WHAT ARE THE IMPLICATIONS FOR FISCAL POLICY?<sup>†</sup>

## Consumption, Saving, and Fiscal Policy

By MICHAEL J. BOSKIN\*

When this year's Nobel Laureate, Robert Solow, and Ely Lecturer, Alan Blinder, teamed their impressive talents several years ago to ask, "Does Fiscal Policy Matter?", the answer they gave was a resounding yes. Working in a then sophisticated *neo-Keynesian* IS-LM tradition, Blinder and Solow presented a parsimonious macroeconomic model and some statistical and historical evidence suggesting that changes in the weighted standardized government surplus did indeed substantially affect real economic activity. The profession was pretty much convinced that the permanent income hypothesis (PIH) or life cycle hypothesis (LCH) *almost* provided a *sufficient* framework for analyzing consumption and saving and hence the effects of fiscal policy. Curiously, the large Keynesian effect of a tax-induced rise in current disposable income was reconciled with the very small effect predicted by the longer-horizon models with more of a whimper than a bang.

Since then, the profession has moved some distance from complete acceptance of the life cycle and permanent income hypotheses and perhaps even further from the traditional consumption function specification in vogue at that time. Indeed, Robert Barro (1974) rekindled the notion that a tax-for-debt swap would have no real effects. An avalanche of analytical and empirical research has sharpened our understanding of

the issues involved, the econometric difficulties in estimating the relevant parameters, and even the care necessary in defining what is meant when one asks whether fiscal policy has any real economic effects. Surprisingly, despite numerous caveats, new and improved data and estimation techniques, and improved perspectives offered by analytical insights not yet prevalent when Blinder and Solow wrote their paper, my conclusion is that their answer is essentially correct: fiscal policy does matter, both for short-run stabilization purposes and for long-run capital accumulation. I believe the preponderance of the evidence strongly supports this view, although the empirical research suggests that the impact of, say, tax cuts on consumption is perhaps *only one-third as large* as the typical Keynesian estimate of two decades ago, but much larger than the neutrality predicted by Ricardian equivalence or the very small effect predicted by the PIH or LCH.

### I. Analytical Issues

A number of theoretical studies cast considerable doubt on the notion that government financial policy is irrelevant. Numerous assumptions necessary to generate this result turn out on close inspection to be unappealing. The failure of Ricardian equivalence when one relaxes the strong assumptions does not tell us very much about how far we deviate from neutrality. But the equivalence of debt and taxes, given government spending, is a striking result. Barro's modern restatement of the proposition is a much more fundamental attack than those dealt with by Blinder and Solow (essentially the shape of IS and LM curves and various measurement

<sup>†</sup>*Discussants:* John B. Shoven, Stanford University; David A. Wise, Harvard University; Angus S. Deaton, Princeton University; Denis Kessler, Fondation pour la Recherche Economique et Financière, Paris.

\*Stanford University, Stanford, CA 94305 and NBER.

problems). In a Ricardian world, a tax cut does not affect aggregate demand.<sup>1</sup>

Careful thought leads one to the conclusion that for government financial policy to matter, it must do something that either the private sector cannot do or will not undo. In short, *it must redistribute resources either among different cohorts or generations; across states of nature in an uncertain world; or among different points in individual households' lifetimes in such a manner as to alter the opportunities available to them.* I believe the most important analytical factors which weigh heavily against pure neutrality include:

1) The absence of a full set of Arrow-Debreu markets. Thus, trades between individuals at different points of time, or between an individual at different points in his life, may not be possible and thus the government may be able to pool risks in a way that the private market cannot (Joseph Stiglitz, forthcoming).

2) Government financial policy will usually generate intergenerational, or inter-cohort, redistribution. This is because the private sector will discount at the sum of the real rate of interest plus mortality probability, whereas the government will discount taxes at the real rate of interest only (Olivier Blanchard, 1985).

3) There is substantial evidence that at least a modest fraction of the population is liquidity constrained at a given point in time (see Robert Hall and Frederic Mishkin, 1982, and my paper with Michael Hurd, 1984). Liquidity constraints raise the marginal propensity to consume out of temporary tax changes to a large multiple of the small amount predicted under perfect capital markets (R. Glenn Hubbard and Kenneth Judd, 1986).

4) Distortionary taxes and uncertainty over future tax burdens or income streams in general imply that financial policy may not

be neutral (Willem Buiter and James Tobin, 1979, and Martin Feldstein, 1982).

5) Widespread intermarriage and parent-child linkages would place virtually everyone into a large interconnected network in which consumption depended only on aggregate resources. Any change in aggregate resources would be divided among a large number of people and therefore, any increment in bequests would be divided among a large number of persons and this would allow only a negligible change in consumption; correspondingly, the potential estator would prefer to leave no bequests at all, and therefore in equilibrium, donors would be driven to corners so no large interconnected network would emerge (Douglas Bernheim and Kyle Bagwell, 1988).

6) Our most fundamental modeling of consumer behavior embodies *very strong assumptions*. Much recent research models aggregate consumption with a representative agent, maximizing an intertemporally separable and stationary utility function, subject to a budget constraint over a long time horizon. One might object fundamentally to utility maximization, extreme rationality, and farsightedness, the computation demands placed on individuals, etc. Myopia, rules-of-thumb, self-control, habit, etc., all may be important, at least for part of the population.

7) Almost all of these discussions of the effects of government financial policy use models of a closed economy. Perhaps the most important consideration for the efficacy of short-run fiscal policy is the openness of the economy to trade and capital flows. While domestic consumption may be stimulated by a tax cut, net exports will decline, thereby offsetting some of the stimulative effect on domestic spending. In the long run, the pattern of capital formation may well be altered.

Of course, demonstrating conditions under which Ricardian equivalence may not, need not, or will not hold *tells us little about how far from* the Ricardian equivalence prediction of *no real effect* of a tax-for-debt substitution we really might be in any particular circumstance. That is, of course, an empirical question.

<sup>1</sup>I take it that everyone believes that an increase—at least an unanticipated increase—in real government purchases raises interest rates, causing a postponement of consumption and increased short-run supply.

## II. Recent Empirical Research on Consumption, Saving, and Fiscal Policy

### A. Traditional Time-Series Studies

There has been a substantial explosion of research along what might be termed extension of traditional time-series aggregate consumption function estimation. The basic approach has been an attempt to expand the traditional set of factors thought to influence aggregate consumption in a given period, to derive better measures of, or proxies for, the relevant variables such as permanent income, wealth, the government budget deficit, or other fiscal and monetary variables.

Consider the following simple specification of private consumption behavior:<sup>2</sup>

$$(1) \quad C_t = \alpha_0 \alpha_1 Y_t + \alpha_2 W_t + \alpha_3 GE_t + \alpha_4 TR_t \\ + \alpha_5 r_t D_t + \alpha_6 T_t + \alpha_7 D_t + \alpha_8 GK_t \\ + \alpha_9 DR_t + \alpha_{10} GKR_t + \alpha Z_t + \varepsilon_t$$

where  $C_t$  is private consumption,<sup>3</sup>  $Y_t$  is national income,  $W_t$  private wealth,  $GE_t$  government expenditure on goods and services,  $TR_t$  government transfers,  $r_t D_t$  government (nominal) interest payments,  $T_t$  taxes,  $D_t$  government net financial debt,  $GK_t$  government nonfinancial capital,  $DR_t$  real revaluations of  $D_t$ ,  $GKR_t$  real revaluations of  $GK_t$ , and  $Z_t$  a vector of other exogenous (for example, demographic) variables.<sup>4</sup>

Note first that traditional consumption function estimates are considerably less general than (1). For example, disposable income,  $Y_t - T_t$ , is often used as the income variable and the government surplus,  $(T_t - GE_t - TR_t - r_t D_t)$ , is aggregated from its components. Referring to (1), this specifica-

tion imposes certain restrictions on the coefficients as maintained hypotheses (in this case,  $\alpha_1 = -\alpha_6$  and  $\alpha_3 = \alpha_4 = \alpha_5 = -\alpha_6$ ). For discussions and tests of such restrictions, see Feldstein (1982) and Bernheim (1987).<sup>5</sup> Ignoring government assets and nonfinancial liabilities, a strong Keynesian would believe the (traditional) deficit and income variables would have coefficients which were equal but of opposite sign (i.e., no future tax discounting), and a strong Ricardian would believe that the coefficient on the deficit would be zero (given government spending). If estimated consistently, the difference in the coefficients on income and the deficit would estimate the effects of a change in government financial policy, that is, a tax-for-debt substitution.

The overwhelming bulk of studies, despite their many measurement, specification, and estimation problems, conclude that deficits do indeed affect consumption (see James Barth, George Iden, and Frank Russek, 1984). Among the influential studies which find a government debt effect on consumption quite similar to that of private wealth are Feldstein (1982), John Seater and Robert Mariano (1985), Blinder and Angus Deaton (1985), myself (forthcoming), and Franco Modigliani and Arlie Sterling (1986). Properly interpreted (see Bernheim, 1987), most studies estimate a net deficit for tax substitution effect on consumption of from 20 to 40 cents per dollar, although some studies are much more dubious (for example, see Paul Evans, 1987). Most studies also reject complete myopia and estimate that future taxes are partially anticipated.

Several studies have attempted to refine measures of  $W_t$ ,  $D_t$ , or  $GE_t$ . For example, an entire line of work dates from Feldstein (1974) and includes a measure of expected (gross or net of tax) future Social Security wealth. Despite numerous difficulties, the consensus appears to be about a 25 to 50 cent decrease in private saving per dollar of Social Security wealth.

<sup>5</sup>Recall the development of weighted standardized deficits in the late 1960's by Gramlich, Hansen, and others. See Blinder and Solow (1974).

<sup>2</sup>Conditions for aggregation and other nontrivial specification problems are ignored.

<sup>3</sup>Proper treatment of consumer durables may well include the imputed rent to the stock in consumption and income, and purchases as investment (see my forthcoming paper with Robinson and Huber for recent estimates).

<sup>4</sup>I ignore variables such as money balances to focus attention on deficit and debt measures.

Several authors have used extended or adjusted measures of  $D_t$  (Robert Eisner and Paul Pieper, 1984; Eisner, 1986; and myself, forthcoming). Eisner and Pieper use an inflation and interest rate adjusted (high employment) deficit. My paper develops separate estimates of government consumption and investment, government tangible capital, and other government assets and liabilities. I estimated the impact of tax cuts on consumption at about 30 cents per dollar; I also estimated that the private propensity to consume out of government tangible assets is similar to that out of ordinary private wealth, suggesting public and private saving are substitutes. While these and related measurement issues for more comprehensive government capital accounting (as discussed in my forthcoming paper with Marc Robinson, and Alan Huber) certainly must be important to measures of national wealth and the legacy left to the future in terms of government service net of future tax liabilities, it is unclear they should replace traditional deficit measures for short-run fiscal policy evaluation. In particular, the appropriate measure of fiscal stimulus depends upon the model of the economy.

A second type of study uses the estimated Euler equation derived from the first-order conditions for optimal consumption behavior under uncertainty, which suggests that consumption should evolve as a random walk or that changes in consumption should not be predictable. These studies estimate the parameters of a stochastic difference equation for consumption, in which the influence of wealth and income on consumption should be zero. The basic question is often interpreted as whether there is an "excess sensitivity" of consumption to income which cannot be explained by people fully rationally optimizing over a long time horizon. Important papers in this tradition date from Hall (1978) and include Marjorie Flavin (1984), Fumio Hayashi (1982, 1985), and the microeconomic data exploration of Hall-Mishkin, which concluded that about four-fifths of consumers could be modeled as if they were maximizing over a long time horizon, whereas one-fifth could not. The traditional interpretation of these results is a test

for liquidity constraints, although other explanations are also plausible.

Thus, 20 percent of the population exhibits Keynesian behavior—they consume out of current disposable income. Clearly, fiscal policy will have some effect in such circumstances. Direct examinations of fiscal policy variables in this context have been less prevalent than the traditional time-series consumption function framework. Roger Kormendi (1983) obtains similar results to those reported above. David Aschauer (1985) finds lagged deficits affect the change in consumption.

The Euler equation approach involves several important advances, especially the ability to circumvent the thorny issue of measuring permanent income. However, most studies assume rather strict maintained hypotheses, for example, maximization by a representative consumer of a utility function (usually taken to be intertemporally separable and stationary), and that the econometrician can specify the information set available to consumers at each point in time. Econometric tests reject these specifications. The empirical usefulness of such aggregate time-series studies is especially clouded by the aggregation issue, the failure to account for real income growth in the relationship between successive observations of aggregate consumption, that relationship depending upon capital income taxation, and the "innovation" in information changing the subjective distribution of expected future government spending, taxes, transfers, and borrowing in a way that may be based (to take an extreme case) on the entire previous history of these variables, not simply subsumed in lagged consumption.

#### B. *Dissaving of the Elderly?*

The strict form of the life cycle hypothesis suggests that the elderly should be dissaving. Numerous studies on cross-section data imply that the elderly seem not to dissave, and in fact, may continue to save. This empirical finding has questioned the applicability of the strict form of the life cycle hypothesis. Related studies attempting to explain consumption and earnings paths of households

to see if aggregate saving can account for a substantial fraction of the capital stock held by households also typically reveal that there is a large unexplained residual (see Laurence Kotlikoff and Lawrence Summers, 1981).<sup>6</sup>

Several more recent studies on longitudinal data (Bernheim, 1984; Peter Diamond and Jerry Hausman, 1984, and Hurd, 1987) find that the elderly do dissave after retirement, although the extent of dissaving is not large in all cases. Hurd also finds no evidence for a bequest motive via differential saving of those who have living heirs.

Thus, I tentatively conclude that the elderly dissave, but the typical individual does so at a rate that still leaves an expected, if modest, bequest. This immediately returns us to the question of the motives for bequests, altruism or otherwise (see Bernheim, 1987, for a discussion).

### C. *Direct Tests of the Relevance of the Age Distribution of Resources*

A more direct test, less susceptible to some of the criticisms of the traditional time-series consumption functions, is performed by myself and Kotlikoff (1985). We build a finite approximation to an intergenerationally altruistic infinitely lived optimal consumption program and test whether the age distribution of resources affects consumption. One of the striking implications of the Ricardian equivalence hypothesis is that the age distribution of resources should not affect aggregate consumption, given the aggregate level of resources. Kotlikoff and I reject this implication of the Ricardian equivalence hypothesis based on postwar U.S. time-series data.

My paper with Lawrence Lau (1988) developed age-cohort-specific balance sheets by combining *Current Population Survey* data on the age distribution of income with the usual aggregate variables. In an econometric

model which satisfies the very weak conditions for demand functions of no money illusion and budget constraints holding, we estimated an economically important and statistically significant effect of the age distribution of human and nonhuman wealth on the share of aggregate wealth consumed. We also estimated a large, and statistically significant, generation effect. Households headed by persons born since 1939 consume a larger share of their wealth than those born prior to 1939, at the same age, and given other variables affecting consumption. This vintage effect is large enough to explain the bulk of the secular decline in the private saving rate in the United States. Again, the strong implication of Ricardian equivalence is generally rejected.

### III. Conclusion

I have presented above a variety of theoretical and empirical reasons to suspect that fiscal policy does indeed matter in the short and long run. The distinction between permanent and temporary changes, anticipated and unanticipated changes, and considerations of discounting future tax liabilities, liquidity constraints, and a variety of other issues have now been embedded in analytical and econometric research.

These empirical results combined with the numerous theoretical caveats to the neutrality proposition suggest that Ricardian equivalence will not hold exactly. The statistical results suggest that, in fact, fiscal policy does indeed affect consumption. The stimulative impact of tax cuts on spending is considerably less than the traditional Keynesian marginal propensity to consume out of current disposable income of 0.75, but considerably more than the zero predicted by the Ricardian equivalence or the 0.05 that would be expected from unconstrained rational intertemporal optimizing consumer behavior.

This impressive array of theoretical and empirical results suggests that government financial policy can indeed affect the real economy. It does not, however, say much about the ability of our political process to design and implement countercyclical fiscal policies. It suggests that fiscal policy is not

<sup>6</sup>The latter study contained a mathematical error which when corrected would increase the fraction of the capital stock which can be accounted for by life cycle saving from 20 to 50 percent—still far less than the total.

impotent; it does not imply that it will be wisely used, given recognition and implementation lags and the numerous other objectives and conflicts in the political process determining changes in fiscal instruments.

## REFERENCES

- Aschauer, David A., "Fiscal Policy and Aggregate Demand," *American Economic Review*, March 1985, 75, 117-27.
- Barro, Robert, "Are Government Bonds Net Wealth?," *Journal of Political Economy*, December 1974, 82, 1095-117.
- Barth, James R., Iden, George and Russek, Frank S., "Do Federal Deficits Really Matter?," *Contemporary Policy Issues*, Fall 1984-85, 3, 79-95.
- Bernheim, Douglas, "Dissaving After Retirement," NBER Working Paper No. 1409, July 1984.
- \_\_\_\_\_, "Ricardian Equivalence: An Evaluation of Theory and Evidence," *Macroeconomic Annual*, 1987, Cambridge: NBER, 1987.
- \_\_\_\_\_, and Bagwell, Kyle, "Is Everything Neutral?," *Journal of Political Economy*, forthcoming, 1988.
- Blanchard, Olivier J., "Debt, Deficits, and Finite Horizons," *Journal of Political Economy*, April 1985, 93, 223-47.
- Blinder, A. and Deaton, A., "The Time Series Consumption Function Revisited," *Brookings Papers on Economic Activity*, 2:1985, 465-511.
- \_\_\_\_\_, and Solow, R., "Analytical Foundations of Fiscal Policy," in A. Blinder et al., *The Economics of Public Finance*. Washington: The Brookings Institution, 1974.
- Boskin, Michael J., "Federal Government Deficits: Some Myths and Realities," *American Economic Review Proceedings*, May 1982, 72, 296-303.
- \_\_\_\_\_, "Concepts and Measures of Federal Deficits and Debt and Their Impact on Economic Activity," in Kenneth Arrow and Michael Boskin, eds., *Economics of Public Debt*, Macmillan, forthcoming.
- \_\_\_\_\_, and Kotlikoff, Laurence, "Public Debt and U.S. Saving: A New Test of the Neutrality Hypothesis," in Karl Brunner and Allan H. Meltzer, eds., *Carnegie Rochester Conference Series on Public Policy: The "New Monetary Economics," Fiscal Issues and Unemployment*, Summer 1985, Vol. 23, 55-86.
- \_\_\_\_\_, and Lau, Lawrence J., "An Analysis of Postwar U.S. Consumption and Savings Behavior," mimeo. Stanford University, 1988.
- \_\_\_\_\_, Robinson, M. S. and Huber, Alan, "Government Saving, Capital Formation and Wealth in the United States, 1947-1985," in R. Lipsey and H. S. Tice, eds., *The Measurement of Saving, Investment and Wealth*, NBER, University of Chicago Press, forthcoming.
- Buiter, Willem H. and Tobin, James, "Debt Neutrality: A Brief Review of Doctrine and Evidence," in George M. von Furstenberg, ed., *Social Security versus Private Saving*. Cambridge: Ballinger, 1979.
- Diamond, Peter A. and Hausman, Jerry A., "Industrial Retirement and Savings Behavior," *Journal of Public Economics*, February/March 1984, 23, 81-114.
- Eisner, Robert, *How Real Is the Federal Deficit?*, New York: Free Press, 1986.
- \_\_\_\_\_, and Pieper, P. J., "A New View of the Federal Debt and Budget Deficits," *American Economic Review*, March 1984, 74, 11-20.
- Evans, Paul, "Interest Rates and Expected Future Budget Deficits in the United States," *Journal of Political Economy*, February 1987, 95, 34-58.
- Feldstein, Martin, "Government Deficits and Aggregate Demand," *Journal of Monetary Economics*, January 1982, 9, 1-20.
- \_\_\_\_\_, "Social Security, Induced Retirement, and Aggregate Capital Accumulation," *Journal of Political Economy*, September/October 1974, 82, 905-26.
- Flavin, Marjorie A., "The Adjustment to Consumption to Changing Expectations about Future Income," *Journal of Political Economy*, October 1984, 89, 974-1009.
- Hall, Robert E., "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, 1978, 86, 971-87.
- \_\_\_\_\_, and Mishkin, Frederic S., "The Sensi-



- tivity of Consumption to Transitory Income: Estimates from Panel Data on Households," *Econometrica*, March 1982, 50, 461-81.
- Iayashi, Fumio, "The Effect of Liquidity Constraints on Consumption: A Cross-Sectional Analysis," *Quarterly Journal of Economics*, February 1985, 100, 183-206.
- \_\_\_\_\_, "The Permanent Income Hypothesis: Estimation and Testing with Instrumental Variables," *Journal of Political Economy*, October 1982, 90, 895-916.
- Iubbard, R. Glenn and Judd, Kenneth L., "Finite Lifetimes, Borrowing Constraints, and Short-Run Fiscal Policy," mimeo., Northwestern University, 1986.
- Iurd, Michael, "Savings of the Elderly and Desired Bequests," *American Economic Review*, June 1987, 77, 298-312.
- \_\_\_\_\_, and Boskin, Michael J., "The Effect of Social Security on Retirement in the Early 1970s," *Quarterly Journal of Economics*, November 1984, 99, 767-90.
- Kormendi, Roger C., "Government Debt, Government Spending, and Private Sector Behavior," *American Economic Review*, December 1983, 73, 994-1010.
- Kotlikoff, Laurence J. and Summers, Lawrence H., "The Role of Intergenerational Transfers in Aggregate Capital Accumulations," *Journal of Political Economy*, August 1981, 89, 706-32.
- Modigliani, Franco and Sterling, Arlie, "Government Debt, Government Spending, and Private Sector Behavior: Comment," *American Economic Review*, December 1986, 76, 1168-79.
- Seater, John J. and Mariano, Robert S., "New Tests of the Life Cycle and Tax Discounting Hypotheses," *Journal of Monetary Economics*, March 1985, 15, 196-215.
- Stiglitz, Joseph E., "On the Relevance or Irrelevance of Public Financial Policy," in Kenneth Arrow and Michael Boskin, eds., *Economics of Public Debt*, Macmillan, forthcoming.

# Consumption, Computation Mistakes, and Fiscal Policy

By LAURENCE J. KOTLIKOFF, WILLIAM SAMUELSON, AND STEPHEN JOHNSON\*

An understanding of the correct model of intertemporal consumption choice is crucial to evaluating the effects of fiscal policies. The debates over whether deficit policy matters (Martin Feldstein, 1974; Robert Barro, 1974) and, if so, how to measure such policy (Robert Eisner and Paul Pieper, 1985; Kotlikoff, 1986) are fundamentally debates about the correct model of consumption. Unfortunately, distinguishing empirically between different consumption theories is a subtle business that has produced no strong conclusions. One problem confronting many tests of alternative consumption theories is that they require joint and quite specific assumptions about preferences, economic resources, and the consumer's information set that may not be justified. In such cases, what is described as a rejection of a particular model may simply be a rejection of restrictive assumptions placed on the model.

A second problem that is also routinely swept under the rug involves the implicit assumption that consumers optimize perfectly given their preferences and resources, and that they correctly value their resources. In order to explore these more fundamental questions we, have conducted an experiment to determine whether individuals, when placed in a controlled life cycle setting, make consistent and coherent consumption choices, and whether they correctly value their future resources. The experiment provides negative answers to both of these questions; in the experiment subjects made significant and systematic errors in their consumption choice, reflecting, in part, an overdiscounting of future income.

This paper reviews several of the findings from our 1987 working paper, and then discusses their implications for fiscal policy. We

give a brief description in Section I of the experiment. Section II describes inconsistencies and errors in consumption choice and traces them to the overdiscounting of future labor income. Section III presents some regression results also pointing to an undervaluation of future resources. Section IV discusses the implications of these results for viewing fiscal policy and suggests the need for additional experiments as well as consumption models that acknowledge, rather than avoid computation problems.

## I. Description of the Experiment

"Experiment" may be a somewhat misleading description of our exercise, but we use it for lack of a better term. The experiment was implemented by an interactive computer program in which subjects key in consumption choices in response to a series of questions. Forty-nine subjects (MBA students and undergraduates at Boston University) participated in the life cycle simulation. Subjects were asked what consumption choices they would make if they were single, faced no uncertainty, had specified levels of future earnings and current assets, knew their ages of retirement and death, and could borrow and save at a specified interest rate. The experiment differs from many in that participants, although paid \$10 each to participate, were provided no salient rewards for their responses. However, it was strongly and repeatedly emphasized at the beginning of the experiment that subjects do their best to respond to all questions on the basis of what would make them most happy given the situation described.

Most subjects took about an hour and a half to finish; some finished within an hour, and some took as long as two hours. Sixty students completed the questionnaire. However, 11 questionnaires were excluded from the analysis either because they contained

\*Boston University and NBER, and Boston University, Boston MA 02215.

key punch errors or because the subject failed to complete one or more sections. Therefore, the results to be discussed are based, in most cases, upon 49 sets of responses.

The questionnaire's basic economic setting can be summarized as follows. The individual in the experiment has just turned 35 and will live to his 75th birthday on which day he dies (with certainty). In his job he earns an annual salary of \$25,000 until he retires on his 65th birthday—that is, he works for thirty years and is retired for ten. The individual can save or borrow as much money as he wishes at 4 percent interest. Subjects were instructed that in the questionnaire setting there is no inflation, deflation, or taxes, no dependents to support, no current or potential health problems, and no uncertainty about the future. All durable goods are rented by the year. Finally, it was assumed that annual consumption expenditures occur and that labor earnings are received on January 1st of each year and that the individual's birthday is also January 1st.

In the various parts of the questionnaire we changed different aspects of the basic setting. Specifically, we changed the initial age, the initial assets, the earnings profile, the interest rate, and the age of retirement. In two parts of the questionnaire, subjects were asked for consumption choices at each age between the initial age and the age of death. In other parts, the subjects were only asked to choose the level of consumption at the initial age specified. In these single-answer parts, we used four key ages, 35, 46, 55, and 69. Several precisely and several economically identical situations were repeated more than once to permit tests of consistency and proper discounting.

## II. Inconsistencies and Errors in Consumption Choice

Since there is no uncertainty in our experimental setting, individuals should make the same consumption choice when facing the same present value of resources and the same interest rate. We tested this hypothesis by constructing 17 pairs of situations in which subjects face identical economic resources (at a 4 percent interest rate), but a possibly

TABLE 1—SUMMARY OF ERRORS MADE IN PRECISELY AND ECONOMICALLY IDENTICAL SETTINGS

	Average	Median	Type	$d(Erns/Res)$
Age 35				
II-IIIId	-.004	.000	1	.000
IIIc-Vc	-.231	-.250	2	.112
Va-Vb	.255	.250	3	.000
Vb-IIIb	.286	.000	2	-.223
IIIb-Va	-.232	-.200	2	.223
Age 46				
II-IIIId	.141	.000	1	.000
IIIa-IVc	.040	.000	2	-.144
IIIc-IVd	-.107	-.087	2	.066
IVa-Vb	.083	.080	3	.038
IVa-IVb	.015	.037	2	.068
IIIb-Vc	-.104	-.065	2	.268
Age 55				
II-IIIId	.264	.045	1	.185
Va-IIIc	.034	.000	2	-.009
IIIc-Vb	.059	.050	2	.056
Vb-Va	.030	.000	3	.026
Vc-IIIb	.088	-.042	2	-.015
Age 69				
II-IIIId	.198	.042	1	.165

Note: Type 1 = Identical circumstances; Type 2 = Same resources, different *earnings/res*; and Type 3 = Same resources, same *earnings/res*, different *earnings* pattern.

different mix of human and nonhuman wealth.

For all but 3 of the 17 cases the average absolute percentage error (defined as the absolute difference between the consumption choices in settings A and B divided by the consumption choice in setting A) exceeds 20 percent. Clearly, this constitutes strong evidence of substantial consumption mistakes. Moreover, consumption errors are widespread across the subjects. Each of the 49 subjects made at least 2 large consumption mistakes—an error in excess of 20 percent in absolute value. Thirty-seven of the 49 subjects made 5 or more large consumption errors in the 17 cases. Thirty-nine subjects made 1 or more very large errors—errors in excess of 40 percent in absolute value and, of these subjects, 11 made 5 or more very large errors.

Table 1 provides a summary of the results for 17 pairs and an indication that many of the consumption errors are systematic. Symbols like IIIb correspond to parts of the experiment providing the 17 pairs of circumstances. In calculating the average and median percentage errors, we divided the

difference in the two consumption choices by the consumption choice in the first of the two parts of the experiment indicated in the table. The table also displays the change in the human wealth share of total resources between the two pairs of questions as well as the way in which the two pairs of questions differed, if at all.

Consider, for example, the age 35 comparison of Part IIIc with Part Vc. In IIIc, the asset level is \$130,000, while it is \$65,000 in Vc. Since total resources are equal in the two cases, the ratio of the present value of earnings to total resources is greater in Vc. In addition, the timing of labor earnings differs. In IIIc, the earnings path is a constant \$25,000 until retirement. In Vc it is \$20,700 from ages 35 to 44, \$31,000 from ages 45 to 54, and \$42,500 from ages 55 to 64. Taking IIIc as the base, the median percentage change in consumption between IIIc and Vc is negative 25 percent. Of the 30 subjects who answered these 2 questions (Vc was added after some initial experiments were conducted), only 3 had nonnegative errors (i.e., they increased their consumption from IIIc to Vc). Some of the errors are quite sizable; 3 subjects reduced their consumption choice by more than 50 percent although they were in exactly the same economic choice situation.

The age 35 comparison of IIIb with Va also involves an increase in the earnings-resource ratio. Again, the median percentage error is negative; it is negative 20 percent. In this case, 10 of the 49 subjects reduced their consumption by 50 percent or more in switching from the IIIb circumstances to the Va circumstances. The age 35 Vb and IIIb comparison is quite similar; here the earnings to resource ratio falls, and while the median error is zero, the mean is .29, with 12 of 49 errors in excess of positive 50 percent. Overall, in 8 of 10 type-2 cases in which the earnings to resource ratio changes, the average error has the opposite sign of the change in the earnings to resource ratio.

In the age 35 comparison of Va and Vb, the earnings to resources ratio is unchanged. Compared with Vb, earnings in Va occur earlier in the life cycle. Again, there seems to be an undervaluation of future earnings. In

this case the median consumption error in switching to Vb is positive 25 percent, and 20 of 49 subjects increase their consumption by 30 percent or more.

### III. Regression Analysis

To test whether consumption is independent of the mix of resources, we ran regressions of the form:

$$(1) \quad C = \alpha + \sigma_1 A + \sigma_2 E + u,$$

where  $A$  denotes the subject's assets, and  $E$  denotes the present value of her future earnings. Of course, the irrelevance of the mix of resources implies that  $\sigma_1$  should equal  $\sigma_2$ . In addition, if preferences are homothetic,  $\alpha$  equals zero. For each subject we estimated (1) separately at ages 35, 46, and 55 (at age 69, future earnings were zero), using the subject's responses to the multiple consumption choices she made at each of these ages.

The results of these regressions indicate that a significant minority of subjects displayed nonhomothetic consumption behavior. At age 35, the hypothesis of a zero intercept was rejected at the 5 percent significance level in 10 cases (of 49), at age 46 in 24 cases, at age 55 in 4 cases, and at age 69 in 8 cases. The age 46 regressions contained the largest number of observations (16 compared to the next largest number 10). Of the 196 estimated constants ( $49 \times 4$ ), 36 intercepts were significantly positive while only 10 were significantly negative. Thus, for the bulk of nonproportional subjects, the predicted APC falls with income.

Table 2 presents a summary of the distribution of assets and earnings coefficients. In 85 percent of the cases (124 of 147 regressions), the earnings and assets coefficients are both positive as predicted by the life cycle model. The coefficient on assets exceeded that on earnings in slightly more than half of the 147 regressions. In total, 41 of 147 (or 28 percent) of the regressions displayed coefficients that are statistically different from one another at the 5 percent level. In these 41 cases, the coefficient on assets exceed that on earnings 25 times. Finally, there is only a single, insignificant

TABLE 2—TESTS OF THE IMPORTANCE OF THE RESOURCE MIX TO CONSUMPTION<sup>a</sup>

	Age			Total
	35	46	55	
Total	49	49	49	147
$\sigma_1, \sigma_2$ Pos	35	44	45	124
$\sigma_1 > \sigma_2$	36	24	17	77
$\sigma_1, \sigma_2$ sig diff	14	11	16	41
$\sigma_1$ sig $> \sigma_2$	14	6	5	25

<sup>a</sup>Entries are number of regressions.

asset coefficient (which is negative) but 16 negative earnings coefficients, 8 of which are significant. It appears from these results that a significant minority of subjects undervalue earnings relative to assets, while a somewhat smaller minority overvalue earnings.

We repeated regression (1) on pooled data. The assumption of equal valuation of assets and earnings is strongly rejected for the pooled age 35 data, but accepted for the pooled age 46 and pooled age 55 data. In the age 35 pooled regression, the assets coefficient is over 7 times greater than the earnings coefficient. These results may reflect an inability of subjects to discount properly far distant earnings streams; that is, at ages 46 and 55 the future earnings streams extend for a shorter interval than at age 35.

We also checked whether squared powers of assets and earnings and the product of assets and earnings help explain consumption in the pooled data. These additional variables are jointly significant for the age 35 and the age 46 regressions.

Two additional results are worth mentioning in passing. First, pooling the data is very strongly rejected for each of the four ages; that is, there is heterogeneity in individual regression coefficients. Second, for a large percentage of subjects the standard time-separable homothetic consumption model explains only a modest fraction of the total variance in consumption choice. For example, if we constrain  $\alpha$  to equal zero and  $\sigma_1$  to equal  $\sigma_2$  in (1), 70 percent of the  $R^2$ s are below .5 for the age 35 subject-specific regressions. The corresponding percentages at ages 46, 55, and 69 are 51, 68, and 23 percent. Hence, assuming the standard

time-separable homothetic life cycle model is correct, error in decision making accounts for a significant fraction of the variance in consumption.

#### IV. Implications for Fiscal Policy, and the Need for More Experiments and Models of Computation Difficulties

The results presented here suggest both an undervaluation of future resources by a significant minority of subjects and an overvaluation by a small minority. An obvious implication of an undervaluation of future relative to current resources is that Ricardian policies that change the timing, but not the present value of taxes will not be neutral. For example, a cut in taxes today coupled with an equal present value increase in taxes in the future will lead to more current consumption. This result is predicted by the *ad hoc* Keynesian consumption model, that places much greater weight on cash flows.

Our findings, however, should be viewed cautiously. First, they are based on only a small number of subjects. Second, although some analysis reported in our study suggests the ability of subjects to abstract from their own circumstances in making consumption choices, more research on this issue is needed. Indeed, considerably more experimental analysis of this kind is needed both to confirm our results and to pinpoint better the source of consumption mistakes.

Also needed are models that consider the costs of computation and bounded rationality. Models of computation costs could take the form of agents making mean zero errors in choosing their consumption, but being able to reduce the variance of these errors at a utility or monetary cost. A different approach would be to view costly computation as involving the acquisition of new information at a cost. The greater the cost of information acquisition, the less information will be obtained and the less calculation will occur. One problem with models of this kind is that they are not likely to predict the kind of systematic mistakes displayed in Table 1. In addition, they assume that it is costly to engage in some, but not all aspects of the decision-making process; for example,

it is costly to gather information, but is not costly to process information. Thinking about the costs of processing information—the costs of thinking—leads to the conundrum that thinking about thinking should itself be costly and the possible implication that rational choice may be individual-specific. If such is the case and individual-specific *ad hoc* approaches really do characterize behavior (Ken Binmore, 1987), it seems that careful controlled experiments may be the best way to start understanding that behavior.

#### REFERENCES

- Barro, Robert, "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, 82, 1095–117.
- Binmore, Ken, "Remodeled Rational Players," mimeo., London School of Economics, 1987.
- Eisner, Robert and Pieper, Paul, "How to Make Sense of the Deficit," *Public Interest*, Winter 1985, 78, 101–118.
- Feldstein, Martin, "Social Security, Induced Retirement, and Aggregated Capital Accumulation," *Journal of Political Economy*, September/October 1974, 82, 905–26.
- Johnson, Stephen, Kotlikoff, Laurence J. and Samuelson, William, "Can People Compute? An Experimental Test of the Life Cycle Model," NBER Working Paper No. 2183, March 1987.
- Kotlikoff, Laurence J., "Deficit Delusion," *Public Interest*, Summer 1986, 84, 53–65.

# Are Consumers Forward Looking? Evidence from Fiscal Experiments

By JAMES M. POTERBA\*

How changes in current and future income affect consumer spending is a perennial question in macroeconomics and public finance. Theoretical constructs such as the permanent income hypothesis are of limited assistance in resolving this issue, because they neglect borrowing constraints and other market imperfections that can significantly affect the marginal propensity to spend out of current income. Recent empirical work has also proven inconclusive, since whether consumption responds to income fluctuations by more or less than the permanent income hypothesis predicts turns critically upon whether disposable income follows a random walk or is stationary around a long-run trend (see John Campbell and Angus Deaton, 1987). This controversy is unlikely to be resolved conclusively, because it is notoriously difficult to distinguish a stationary but highly persistent time-series from one that is nonstationary. These difficulties suggest the importance of searching for "natural experiments," income shocks with predictable and well-understood effects on future income, to test models of consumer behavior. Changes in federal tax and transfer policy during the last two decades provide several episodes of this type.<sup>1</sup> Analyzing

the effects of these tax changes is also central to understanding the role of fiscal policy in affecting national saving.

This paper examines two aspects of consumer response to tax changes. Section I argues that consumption responds to temporary income tax shocks by more than the permanent income hypothesis would suggest. Results from the 1975 tax rebate in particular suggest that a \$1 increase in transitory income raises spending by about 20 cents. Section II examines consumption responses to tax announcements. Although the results are not conclusive, they suggest that some consumers do not adjust consumption in anticipation of tax changes. The final section sketches some implications for analyzing fiscal policy.

## I. Consumption Changes Induced by Temporary Tax Policies

There have been two explicitly temporary federal income tax policies in the last two decades, the 1975 income tax rebate and the 1968 surtax.<sup>2</sup> The Tax Reduction Act of 1975 consisted of a 10 percent rebate of 1974 taxes up to a maximum of \$200. The House passed the bill in late February, the Senate approved it in March, and President Ford signed it in late March. The legislation transferred \$8.1 billion from the Treasury to households, mostly during the two months from late April to mid-June. Measured at an annual rate in 1987 prices, this corresponds to a disposable income increase of more than

\*Department of Economics, MIT, Cambridge, MA 02139. I am grateful to the NSF for research support, and to Angus Deaton, Greg Mankiw, David Romer, John Shoven, Lawrence Summers, and David Wilcox for helpful discussions. A data appendix is on file with the ICPSR in Ann Arbor, Michigan.

<sup>1</sup>Tax changes are attractive fiscal experiments because of their simple stochastic structure, but they also have drawbacks. The neoclassical view of fiscal policy (see Robert Barro, 1987) holds that changes in taxation, unlike other changes in disposable income, should not affect consumption. Lawrence Summers and I (1987) describe the apparent failure of this view to account for the recent coincidence of high fiscal deficits and depressed personal saving.

<sup>2</sup>Alan Blinder and Deaton (1985) note that the income tax reductions in the Economic Recovery Tax Act of 1981 can also be viewed as a permanent tax reduction, accompanied by two years of temporary tax *increases*. Their results, however, suggest that consumers viewed the tax change as a sequence of permanent tax cuts.

\$100 billion. By comparison, the 1968 surtax raised taxes by \$16 billion and the pre-announced 10 percent income tax reduction of 1982 lowered taxes by \$31.6 billion (\$1987). The 1975 tax bill included the tax rebate as well as a smaller, transitory income tax reduction that was subsequently made permanent; see Blinder (1981). There was also a one-time Social Security bonus for retired individuals with no taxes to rebate. My analysis focuses on the rebate's consumption effects, since it is difficult to describe consumer expectations with respect to the other tax changes.

Studies of the 1975 rebate as well as earlier work on the 1968 surtax surveyed in Blinder yield conflicting evidence. While Blinder concluded that each rebate dollar raised consumption by about 16 cents in the quarter when it was received, he also found substantial effects between five and eight quarters after the rebate. These estimates are larger than the marginal propensity to consume out of windfalls under the permanent income hypothesis, but not as large as his estimates of the propensity to consume from a permanent tax reduction. Franco Modigliani and Charles Steindel (1977) found much smaller effects from the 1975 rebate. Blinder and Deaton estimated a more complex consumption specification, and in part because of their large standard errors, they could reject neither the view that consumers respond only to current income nor the view that they ignored the rebate completely.

My analysis differs from previous studies in using monthly consumption data to exploit the pronounced intraquarter pattern of the 1975 rebate. (More than three-quarters of the rebate checks were received in May.) This higher frequency data is also attractive because finding that spending rises in the month when tax payments change is strong evidence of a link between current income and consumption.

I estimate the effect of the rebate and surtax following recent econometric studies of the stochastic permanent income hypothesis (see Robert Hall, 1987). These studies demonstrate that in the presence of rational expectations and perfect credit markets, the change in the logarithm of real per capita

consumption ( $c_t$ ) should not be predictable from lagged information. Since both the rebate and surtax were announced at least a month before the change in tax payments, they should have been incorporated into consumption before they affected cash flow. The tax changes therefore should not help forecast the change in consumption between one month and the next. I test this by estimating

$$(1) \quad c_t = \alpha_0 + \alpha_1 c_{t-1} + \gamma [\Delta \text{tax}_t / c_{t-1}] + \varepsilon_t$$

where  $\Delta \text{tax}_t$  denotes the change in the real per capita rebate level in month  $t$ .<sup>3</sup> After scaling  $\Delta \text{tax}_t$  by lagged per capita consumption, the coefficient  $\gamma$  measures the amount of additional consumption that results from a \$1 rebate.

Equation (1) omits the possibility that other news, coincident with the changes in tax policy, might explain shocks to consumption. I therefore also estimate an expanded version of (1), including both current and lagged values of several other variables—real per capita wage and salary income, short-term nominal interest rates, and stock market returns—to control for these factors. I estimate both specifications for the 1959:6–1987:9 sample period (339 observations) using three different measures of consumption: nondurables, services, and the consumption measure developed by Blinder-Deaton. The latter adds nontax payments for government services to consumption of services and nondurables other than shoes and clothing. Analyzing consumption components individually requires that the utility function be additively separable between components, a strong assumption that has been made in many earlier studies.

Table 1 presents estimates of the  $\gamma$  coefficient, with separate estimates for the transitory tax events in 1968 and 1975. Positive

<sup>3</sup>Data for the 1975 surtax are drawn from monthly Treasury data on personal tax payments. The *Survey of Current Business* for May 1978 reports the impact of the 1968 surtax on withheld tax payments; I distribute the quarterly values equally across months.



TABLE 1—CONSUMPTION EFFECTS OF 1975  
TAX REBATE AND 1968 SURTAX

Consumption Measure	Consumption Effect of \$1 Rebate ( $\gamma$ )			
	Basic Specification		Augmented Specification	
	1968	1975	1968	1975
Nondurables	0.045 (0.259)	0.246 (0.102)	0.049 (0.233)	0.182 (0.092)
Services	0.065 (0.129)	0.014 (0.055)	0.087 (0.123)	0.011 (0.052)
Nondurables + Services- Shoes-Clothes + Nontaxes	0.227 (0.278)	0.167 (0.118)	0.275 (0.252)	0.134 (0.107)

Note: Standard errors are shown in parentheses. Specifications are described in the text.

coefficients indicate that consumption moved in the same direction as the tax-induced change in disposable income. The results differ somewhat across specifications, but suggest that consumption spending rises by between 12 and 24 cents per dollar of temporary tax rebate. The specification constrains a temporary tax change to have equal and opposite signed effects on consumption outlays in two consecutive months, and does not allow for additional effects in later months. Some experimentation with additional lagged values of  $\Delta tax$  did not yield any clear results.

Outlays on nondurables appear more sensitive to disposable income fluctuations than do outlays on services. The standard errors for most of the estimates are nevertheless large, and only the results for the increased spending on nondurables associated with the 1975 tax rebate is statistically significant at standard levels. The point estimates are nevertheless quite close to those that Campbell and N. Gregory Mankiw (1987) obtain from structural estimates of a consumption-income system, and those from Blinder's study of quarterly consumption data. These studies together constitute evidence against the view that changes in the timing of taxes do not affect real activity.<sup>4</sup>

<sup>4</sup>Not all tax-induced transitory income appears to affect consumption. David Wilcox (1987a) shows that the 1985 delay in income tax refunds did not affect consumer spending.

## II. Does Consumption Respond to Tax Announcements?

Monthly consumption data can also be used to investigate whether consumers respond to announced, but not yet implemented, tax policies. In addition to the temporary tax changes of Section I, I examine announcements of the permanent personal tax reductions of 1964, 1981, and 1986 as well. The 1964 tax reform reduced personal income taxes by approximately 15 percent. Although President Kennedy first proposed a tax reduction in 1962, Congress passed and President Johnson signed the Act in February 1964. The law's first adjustments to withholding occurred in late March. The 1981 Economic Recovery Tax Act, proposed by President Reagan in February, received congressional approval in July and became law in August. It phased in personal tax reductions totalling 25 percent over the next three years. The 1981 changes were not reflected in tax withholding until mid-October. The 1986 Tax Reform Act received congressional approval in August. It had only minor effects on personal tax liability for 1986, but reduced personal taxes by \$19.2 and \$29.6 billion in 1987 and 1988.

It is extremely difficult to determine when consumers changed their beliefs about the likelihood of various legislative outcomes. George Katona and Eva Mueller (1968) report that one month prior to passage of the 1964 tax bill, 40 percent of individuals did not expect a tax reform to become law. In late 1963, less than one-third expected tax reform. While somewhat arbitrary, my analysis takes the month of congressional passage as focal for expectational changes. For most of the revenue bills considered here, any uncertainty about presidential approval once the bill cleared Congress was resolved quickly. Undoubtedly expectations also evolved in earlier months. The substantive conclusions presented below are not altered, however, by considering consumption changes for several months up to and including congressional passage of the tax laws.

Table 2 reports the residuals for the tax passage months from the expanded version of equation (1), including news about wages

TABLE 2—CONSUMPTION RESPONSE TO  
PASSAGE OF TAX BILLS

Date	Consumption Measure <sup>a</sup>		
	(1)	(2)	(3)
February 1964	-0.329 (0.711)	0.080 (0.366)	-0.115 (0.383)
June 1968	0.857 (0.702)	0.627 (0.359)	0.692 (0.380)
March 1975	-0.369 (0.706)	0.232 (0.360)	0.154 (0.379)
August 1981	-0.691 (0.712)	-0.408 (0.363)	-0.407 (0.385)
August 1986	0.450 (0.712)	0.098 (0.363)	0.201 (0.384)

*Note:* Standard errors are shown in parentheses. Estimates are based on the augmented version of equation (1), including current and lagged values of wages, stock returns, and interest rates. Each entry denotes the given month's residual from the augmented consumption model.

<sup>a</sup> Col. 1 denotes Nondurables; Col. 2 denotes Services; Col. 3 denotes Nondurables + Services + Nontaxes-Shoes-Clothing.

and rates of return as well as lagged consumption. If important news about the present value of future tax liabilities was released in these months, without accompanying information about the future course of spending, then we should observe consumption revisions. The results provide little support for the view that consumers react to the announcement of tax policies. The month in 1968 when the Vietnam War surtax was enacted coincides with a significant increase in consumption. The month when passage of the dramatic 1981 tax cut was finally assured shows a downturn in consumption by all three measures. Parallel findings emerge for the 1964 tax reform. Passage of the 1975 and 1986 bills had the expected effect on consumption, but the point estimates are too small to reject the null hypothesis of no effect.<sup>5</sup>

<sup>5</sup>Albert Jaeger and Klaus Neusser (1987) corroborate these findings with Austrian data on tax changes and consumption, finding virtually no effect for tax announcements.

These results suggest that news of policy changes does not significantly affect consumer spending, even though enactment of the policies does matter. The findings in Table 2 must be viewed as tentative, since the standard error on the monthly consumption change is large. In addition, my "event-study" approach is likely to have relatively low power especially against the alternative hypothesis that consumers gradually revise their expectations of future tax policy and adjust consumption accordingly. Nevertheless, the results may warrant further attempts to measure the real effects of pre-announced fiscal policies.

One related source of evidence on how spending responds to policy announcement and implementation is Wilcox' (1987b) study of Social Security benefit increases between 1965 and 1985. These increases were often substantial: 20 percent in 1972, and more than 10 percent in five other years. After analyzing the effect of Social Security benefit increases on both personal consumption spending and on retail sales, Wilcox rejects the view that these changes in transfers do not affect consumer behavior. He finds that when a \$1 benefit increase takes effect, retail sales increase by approximately \$1.30. About 85 cents of the increased expenditure is on consumer durables, while the remainder is divided between nondurables and services. Spending in excess of the transfer increment may reflect one-time purchases of durables. Even though benefit changes were all announced at least six weeks prior to enactment, there is no evidence that consumption rose when benefit increases were legislated. This supports the earlier analysis of the temporary tax changes, although it is weak support since Social Security recipients may be especially prone to facing binding liquidity constraints.

### III. Conclusions

Two lessons can be drawn from the "fiscal experiments" of the last two decades. First, a transitory tax-induced income increase raises consumer spending by roughly one-fifth as much. This exceeds the amount of consump-

ion that a life cycle or permanent income model would suggest, but it also implies a marginal propensity to consume that is well below the average propensity. Contrary to the Ricardian equivalence view, the timing of taxes appears to affect the level of real activity.

Second, while implementation of tax policies has a detectable effect on consumption, spending responds by relatively little to policy announcements. This may be explicable by some feature of the consumption technology that induces lags between news of future income movements and actual spending. For example, consumers may need to engage in time-consuming search before purchasing some goods. Habits may also be important in consumption choices. Perhaps the simplest explanation, however, is that a significant fraction of consumers are either myopic or face liquidity constraints that prevent them from adjusting consumption in response to news about future disposable income. Myopia, if part of the explanation, is clearly not universal. There are obvious cases such as retiming income realizations across calendar years with different tax rates that suggest substantial numbers of taxpayers are responsive to preannounced changes in policy. The existence of such taxpayers, however, does not disprove the existence of others who fail to look ahead.

The existence of myopic consumers has important implications for fiscal policy. First, theoretical and simulation studies of dynamic tax policy invariably assume that anticipated policies alter current spending patterns. A nontrivial group of myopic individuals necessarily blunts this adjustment, and implies different trajectories for the capital stock and consumption. This issue is of more than academic interest. The Social Security reforms enacted in 1983 but scheduled to take effect between 1997 and 2009 substantially reduce the present value of benefits that today's young workers can expect to receive. The incidence of myopia is a central determinant of these policies' impact on saving. While some have described recent fiscal policy as contractionary on account of these changes, there is little evidence that

these changes have affected consumer behavior.

Second, myopia raises hard questions about the meaning of horizontal equity. Pre-announced tax policies are often promoted because they cause less dramatic changes in the relative well-being of otherwise identical individuals who have historically made different choices. If some consumers fail to recognize the future changes that have been enacted, for example, by failing to save more for their retirement, implementing the preannounced policy will exacerbate inequalities between myopic and farsighted individuals. Whether the presence of some myopic individuals affects the government's ability to commit now to enact substantial policy reforms at some future date is an intriguing question for future work.

## REFERENCES

- Barro, Robert**, "The Neoclassical Approach to Fiscal Policy," in his *Handbook of Modern Business Cycle Theory*, New York: Wiley & Sons, 1987.
- Blinder, Alan S.**, "Temporary Income Taxes and Consumer Spending," *Journal of Political Economy*, February 1981, 89, 26-53.
- and **Deaton, Angus**, "The Time Series Consumption Function Revisited," *Brookings Papers on Economic Activity*, 2:1985, 465-511.
- Campbell, John Y. and Deaton, Angus**, "Is Consumption too Smooth?," NBER Working Paper No. 2134, January 1987.
- Campbell, John Y. and Mankiw, N. Gregory**, "Permanent Income, Current Income, and Consumption," NBER Working Paper No. 2436, November 1987.
- Hall, Robert**, "Consumption," in Robert J. Barro, ed., *Handbook of Modern Business Cycle Theory*, New York: Wiley & Sons, 1987.
- Jaeger, Albert and Neusser, Klaus**, "Excess Consumption as a Predictor of Future Income Changes," mimeo., University of Vienna, 1987.
- Katona, George and Mueller, Eva**, *Consumer Responses to Income Increases*, Washing-

ton: The Brookings Institution, 1968.

**Modigliani, Franco and Steindel, Charles**, "Is a Tax Rebate an Effective Tool for Stabilization Policy?," *Brookings Papers on Economic Activity*, 1:1977, 175-209.

**Poterba, James M. and Summers, Lawrence H.**, "Finite Lifetimes and the Effects of Budget Deficits on National Saving," *Journal of Monetary Economics*, September 1987, 20,

369-92.

**Wilcox, David W.**, (1987a) "Income Tax Refunds and the Timing of Consumption Expenditure," mimeo., Federal Reserve Board of Governors, May 1987.

\_\_\_\_\_, (1987b) "Social Security Benefits, Consumption Expenditure, and the Life Cycle Hypothesis," mimeo., Federal Reserve Board of Governors, May 1987.

# THE ESTIMATION AND MEASUREMENT OF SPILLOVER EFFECTS OF R&D INVESTMENT<sup>†</sup>

## Industry Effects and Appropriability Measures in the Stock Market's Valuation of R&D and Patents

By IAIN COCKBURN AND ZVI GRILICHES\*

Firms have a variety of possible mechanisms for preventing competitors from taking advantage of their investment in knowledge capital, and the availability and effectiveness of these mechanisms varies across firms and industries. In particular, the effectiveness of patents as a mechanism for appropriating the returns from R&D is not a constant, and the present value of returns to a firm from investing in patent protection should differ, therefore, according to industry conditions and firm-specific factors. Failure to take this into account may have been the cause of some puzzling results in earlier work, where patent variables became insignificant (and in some cases wrong-signed) in the presence of R&D variables. Survey data collected by the Yale group (R. Levin et al., 1984) has made it possible, at least in principle, to construct measures of the appropriability of R&D at the industry level. This paper presents results obtained from matching the Yale survey to the NBER data on R&D and patenting intensity of large U.S. manufacturing corporations in an

attempt to control for interfirm variability in the patenting environment.

### I. The Equation to be Estimated

Following Griliches (1981), we estimate the market's relative valuation of tangible and intangible capital. In a rational stock market, a firm's stock price should be the expected discounted value of the net income which will be derived from its assets. Under constant returns to scale, or as a local linear approximation we can write:

$$(1) \quad V = b[A + \delta K]$$

where  $V$  is the market value of the firm,  $b$  is the average multiplier of market value relative to the replacement cost of total assets,  $A$  is tangible capital,  $K$  is intangible capital, and  $\delta$  is its relative shadow price. Dividing through by  $A$ , taking logarithms, and exploiting the fact that  $\log(1+x) \approx x$  when  $x$  is small, we obtain

$$(2) \quad \log(q) \approx \alpha + \delta K/A$$

which is interpretable as a regression equation in which  $K$  is a vector of variables representing a firm's intangible assets. We use "stocks" of R&D and patents built up from annual reported R&D expenditures and numbers of patents granted to proxy for intangible knowledge capital. To the extent that the valuation of such intangible assets (for example, patents) varies from industry to industry, the estimated  $\delta$ s need not be identical. We explore this possibility by interacting the various measures of intangible

<sup>†</sup>Discussants: Adam Jaffe, Harvard University; Peter Reiss, Stanford University; Mark Schankerman, New York University; Wesley Cohen, Carnegie-Mellon University.

\*Department of Economics, Harvard University, Cambridge, MA 02138. We are indebted to the National Science Foundation (PRS85-12758 and SES82-08006), the Sloan Foundation, and the NBER Productivity Research Program for the financial support of this work. We are grateful to Richard Levin for providing us with access to the original Yale survey data, and to Peter Reiss for helpful comments on an earlier version of this paper.

capital with industry level indices of the "ease of appropriability" derived from the Yale survey data.

## II. The Data

We have combined two separate data sets in this study: the NBER RNDPANEL data set and the Yale survey results, which are described more fully in our paper (1987), C. Cummins et al. (1985), and Levin et al. The RNDPANEL is a large data set in panel format compiled from the Compustat and Patent Office tapes. Over 1800 firms are represented, with accounting data and patent figures from the late 1960's through 1984. We use 1980 data for a cross-sectional extract of 722 manufacturing firms, from which we calculate our dependent variable, the log of  $q$ , and our major independent variables:  $K$ , the cumulated "stock" of past  $R\&D$  expenditures (using a 15 percent depreciation rate);  $SP$ , the stock of cumulated past patents (using a 30 percent depreciation rate); and  $NR$ , an estimate of the current year's net investment in  $R\&D$ , which is calculated as  $NR = R\&D - 0.15K$ , where  $R\&D$  is the current year's  $R\&D$  expenditure and  $K$  is the stock of  $R\&D$  carried forward from the end of the previous year. These three variables have all been divided by the total fixed assets of the firm and interacted with the appropriability measures.

The Yale survey took the form of a questionnaire on Industrial Research and Development mailed to  $R\&D$  executives in 1562 business units in over 130 industries defined at the Line of Business level. In all, 650 usable responses were obtained. The questionnaire posed over 120 questions about various mechanisms for appropriating returns from  $R\&D$  and their effectiveness, the nature of technical progress, and the general relevance of science. Respondents were typically asked to answer on a 7-point scale from 1 = *not important* to 7 = *very important*.

In the exploratory stage of this study, we constructed many different appropriability measures from the raw survey responses. We tried nonlinear transformations of scale, means and maxima of various sets of ques-

tions, variables based on the distribution of responses within industries, and also common factors extracted from the correlation matrix of individual responses across questions. The ability of these variables to usefully measure interindustry differences in appropriability can be judged, at least in the first instance, by testing for equality of means across industries. We did an analysis of variance for each of the variables constructed from the survey (which gives a simple  $F$ -test for the equality of industry means) using two definitions of industries devised by us from the SIC product classification (55 "3.5 digit" industries) and the NSF's breakdown of  $R\&D$  data (24 "2.5-digit" industries). For both levels of aggregation, we were able to accept the null hypothesis of equality of industry means at the 5 percent level for all of our nonpatent appropriability measures. It is not obvious that there is much systematic between-industries variance in these variables, and hence their quality as indicators of interindustry differences in the appropriability environment may be rather low. On the other hand, variables based on questions about the effectiveness of patents showed a significant difference in industry means. In all cases, however, excluding from the sample the two extreme industries, drugs and computers, from the sample brought about a marked decrease in the  $F$ -statistic.

Among the patent-based measures, a simple sum of the responses to questions IA1 (do process patents prevent duplication?) and IB1 (do product patents prevent duplication?), which we call *PPP* (Patents Provide Protection) did as well or better than any of the more complicated variables we constructed, and this is the main appropriability measure we use in the rest of our analysis.

The survey respondents were asked to answer questions based on their assessment for their industry as a whole, not for their firm, therefore we interpret responses from individuals in a particular industry as repeated measurements of the industry-level parameter, not as observations of firm-specific effects. Taking the industry mean of individual respondents' scores for an appropriability measure within the survey data

set gave us an estimate of the industry-level parameter, which we then matched to each of our firms in the same industry.

### III. Estimation Results Controlling for Appropriability

Table 1 presents most of our major results. In the absence of *R&D* variables, past patenting does appear to capture some relevant aspects of "intangible" capital. Its coefficient is statistically significant and implies a valuation of approximately \$0.5 million per patent granted. This is consistent with other evidence assembled in Griliches, A. Pakes, and B. Hall (1987). However, when measures of *R&D* are added to the equation this estimate either disappears (col. 2) or is heavily attenuated. Adding a measure of the effectiveness of patent protection to these equations and interacting it with the patent stock and *R&D* variables improves the fit only marginally (by about .01) but does indicate the presence of an interaction. Without *R&D* variables the results imply a much higher valuation of patents in industries where patent protection is more effective. For example, column 4 could be read as indicating an average value of a patent of about \$0.4 million which rises to about \$1.0 million per patent in industries where the effectiveness of patents is two standard deviations higher than the average. When *R&D* variables are added in, the patent stock variables become less significant and the interaction is now attached to the *R&D* stock or the *R&D* "news" variable. Column 6 implies that the market values "news" in *R&D* much more highly than past investments or old patents and that such new *R&D* moves are valued about 50 percent higher in industries where patent protection is more likely to be effective. Adding separate industry intercepts to these equations attenuates these results somewhat, but does not eliminate them entirely (compare cols. 3 and 5).

In regressions not reported here, we tested other appropriability measures: transforming *PPP* by changes of scale, or measuring it by the excess of "high" responses over "low" responses within an industry makes surpris-

TABLE 1—THE STOCK MARKET'S RELATIVE VALUATION OF R&D AND PATENTS

	1	2	3
<i>SP/A</i>	0.493 (0.165)	0.111 (0.094)	0.192 (0.158)
<i>PPP</i>			-.012 (0.017)
<i>PPP*SP/A</i>			0.076 (0.099)
<i>K/A</i>		1.374 (0.182)	1.442 (0.174)
<i>PPP*K/A</i>			0.303 (0.115)
$\bar{R}^2$	0.027	0.125	0.133
	4	5	6
<i>SP/A</i>	0.380 (0.171)	0.107 (0.167)	0.249 (0.155)
<i>PPP</i>	0.034 (0.024)	0.019 (0.024)	0.019 (0.023)
<i>PPP*SP/A</i>	0.236 (0.116)	0.075 (0.110)	0.098 (0.101)
<i>K/A</i>		0.932 (0.201)	0.335 (0.178)
<i>PPP*K/A</i>		0.365 (0.130)	
<i>NR/A</i>			11.96 (1.368)
<i>PPP*NR/A</i>			2.788 (1.231)
$\bar{R}^2$	0.172	0.200	0.310

Notes: Dependent variable:  $\text{Log}(q)$ . Equations 4–6 also contain 10 2-digit industry dummies.

$N = 722$ . Mean of the dependent variable =  $-0.272$ , standard deviation =  $0.697$ . Heteroscedasticity-consistent standard errors are shown in parentheses. All equations also contain an intercept term and the logarithm of Assets, whose coefficient was small but consistently significant, on the order of  $-0.03$  (0.01).

ingly little difference to the results, either in terms of the overall fit of the regressions, or the significance of particular coefficients. The variables intended to measure the effectiveness of nonpatent appropriability methods, which had already been cast into doubt by the ANOVA results, performed badly on their own, and showed no ability to increase the proportion of explained variance or produce significant interaction terms when used in addition to the patent-based measures.

We also combined our 1980 cross section with 1973 and 1979 data in a SUR framework to take account of left-out individual

firm effects. This improved our results marginally, in the sense of a slightly better fit and slightly smaller standard errors, but the overall conclusions remained the same. The method by which we constructed the appropriability measures allowed us to estimate the variance of the measurement error (the sampling error of the industry means), making it possible to correct for errors-in-variables by "de-attenuating" the cross-products matrices in our regressions (see W. Fuller, 1980). This exercise was marginally successful for the simple equations, but failed for more complex equations due to the high proportion of measurement error relative to the total variance of the appropriability variables. We tried a different approach to the errors-in-variables problem by using another patent-based appropriability measure as an instrument for *PPP*. Again, we saw some improvement in the results, but no change in their overall flavor.

There remains the question how much we gain by using such measures of the effectiveness of appropriability mechanisms compared to a cruder interaction with 2-digit level industry dummies. If instead of interacting the *SP/A* and *NR/A* variables with *PPP*, we interact them with our 10 industry dummies, we get quite small increases in the adjusted  $R^2$ s (for example, .315 vs. .310 for the equation in col. 6.) In this sense the *PPP* variable does quite well. It effectively accomplishes the same thing as 10 dummy variable cross-product terms and because it uses up only one degree of freedom it provides a more powerful test of the underlying hypothesis and a more useful interpretation of the data.

#### IV. Conclusions

We tried to improve upon our estimates of the stock market's valuation of knowledge capital embodied in *R&D* and patents stocks by bringing in measures of the appropriability environment facing a firm from the Yale survey. We found the responses to the questions about the effectiveness of patents as a mechanism for protecting the returns from innovation to be of some use. There is some

evidence of an interaction between industry level measures of the effectiveness of patents and the market's valuation of a firm's past *R&D* and patenting performance, as well as its current *R&D* moves. There is no evidence, however, that other appropriability mechanisms differ enough across industries to leave measurable traces in such data. Because the within-industries variance of the survey responses is so high, even for the somewhat better-defined patents questions, our estimates are not very stable, and attempts to improve upon them using various errors-in-variables "de-attenuation" and instrumental variables methods were not particularly successful. Nevertheless, while these effects are not precisely estimated, they are not small. The estimated changes in  $q$  for the average firm implied by an increase of two standard deviations in the effectiveness of patents range from 10 to over 25 percent, which is a rather large effect indeed. Given that *R&D* capital is about 14 percent of the value of all other assets, this implies that such a change in the appropriability environment would come close to doubling its valuation.

The basic message of this paper is consistent with earlier work. There is some interesting information in patent counts, but it is subject to much error. Data on *R&D* expenditures, where available, are stronger measures of input to the process by which firms produce technical innovation than patents are of its "output." This difficulty with the patents numbers is not really eased by adding industry-level information on their relative effectiveness as a means of securing returns from innovation. But appropriability measures do appear to matter: we find significant interactions with either the patent stock or the *R&D* stock variables, implying that the market recognizes that similar *R&D* moves may have different payoffs in different appropriability environments. An alternative interpretation, which needs to be explored further, is that different appropriability environments imply different depreciation rates for *R&D* investment. These should have been incorporated in the construction of the *R&D* "capital" stock and the estimated interac-



tions are our attempt to adjust for not having done so. We shall pursue some of these leads in our future work in this area.

#### REFERENCES

- Cockburn, I. and Griliches, Z., "Industry Effects and Appropriability Measures in the Stock Market's Valuation of R&D and Patents," NBER Working Paper No. 2465, 1987.
- Cummins, C., et al., "The R&D Masterfile: Documentation," unpublished, NBER, 1985.
- Fuller, W., "Properties of Some Estimators for the Errors-in-Variables Model," *Annals of Statistics*, 1980, 8, 407-422.
- Griliches, Z., "Market Value, R&D and Patents," *Economics Letters*, 1981, 7, 183-87.
- \_\_\_\_\_, Pakes, A. and Hall, B., "The Value of Patents as Indicators of Inventive Activity," in Partha Dasgupta and Paul Stoneman, eds., *Economic Policy and Technological Performance*, Cambridge: Cambridge University Press, 1987.
- Levin, R. et al., "Survey Research on R&D Appropriability and Technological Opportunity," research paper, Yale University, 1984.

# Appropriability, *R&D* Spending, and Technological Performance

By RICHARD C. LEVIN\*

A quarter-century ago, Kenneth Arrow (1962) drew the attention of economists concerned with technological change to the consequences of positive externalities associated with private investment in industrial research and development (*R&D*). He observed that a firm's incentive to invest in *R&D* is attenuated when the knowledge generated by the investment is involuntarily transmitted to competitors. Arrow indicated that market economies tend to resolve this problem by the assignment of intellectual property rights, but he stated quite clearly that "no amount of legal protection can make a thoroughly appropriable commodity of something so intangible as information" (p. 615). He noted in particular that leakages of technological knowledge were inevitable, given the embodiment of knowledge in products and the mobility of personnel among firms. In such a regime, Arrow concluded, underinvestment in *R&D* was likely, but he also remarked that "from the standpoint of efficiently distributing an existing stock of information, the difficulties of appropriation are an advantage" (p. 616).

Arrow's observation that returns from *R&D* are neither fully appropriated nor entirely dissipated was largely overlooked in the theoretical literature of the next two decades. Despite a few notable papers on the effects of rivalry between innovators and imitators, much formal analysis of *R&D* investment proceeded by assuming, as Arrow did in the model presented at the conclusion of his classic paper, that the knowledge generated by *R&D* is perfectly appropriable. Even the literature on optimal patent lifetimes, that treated explicitly the welfare

tradeoff between the efficient creation and dissemination of knowledge, characterized the degree of protection afforded by a patent as perfect (or at least constant) over the patent's life and nil thereafter.

## I. Spillovers, Incentives, and Performance in Theory

Theoretical interest in the effect of imperfect *R&D* appropriability on economic performance was revived recently by Michael Spence (1984). In the context of a simple, effectively static, model, Spence found that an increase in spillovers (parameterized as the fraction of a firm's *R&D* that is effectively utilized by its competitors) reduces the incentive to invest in *R&D*, but it also reduces the *R&D* required to achieve a given level of cost reduction. In Spence's model the net effect on *R&D* spending is unambiguously negative, but Spence provided an example in which an increase in spillovers (a decrease in appropriability) increased total economic surplus.

In a model that generalizes Spence's cost function and treats the number of firms as endogenous, Peter Reiss and I (1986) have shown that an increase in Spence's spillover parameter reduces industry *R&D* intensity and increases the equilibrium number of firms. The effect of intra-industry spillovers on technological performance, measured as the inverse of unit cost, depends in general on the elasticity of demand, the elasticity of unit cost with respect to own and borrowed *R&D*, and the level of the spillover parameter. Over a wide range of parameter values, an increase in spillovers improves technological performance (reduces unit cost).

## II. Recent Empirical Work

Although the measurement of intra-industry spillovers in *R&D* has been the sub-

\*Professor of Economics, Yale University, New Haven, CT 06520-1972. The research discussed in this paper was supported by the Division of Policy Research and Analysis of the NSF. Alvin Klevorick offered helpful suggestions, and Somi Seong provided valuable research assistance.

ect of recent attention (see Edwin Mansfield et al., 1981, and Mansfield, 1985), most attempts at measurement have stopped short of linking the extent of spillovers to the magnitude of R&D investment and to technological performance. Evidence suggesting a linkage between spillovers and performance was reported by Adam Jaffe (1986), who assigned firms to technological clusters on the basis of patenting patterns and found that firms in clusters representing large 'pools' of R&D tended to engage in more patenting. Jeffrey Bernstein and M. Ishaq Nadiri (1986), who estimated systems of dynamic factor demand equations in four industries, found that an increase in intra-industry R&D spillovers decreased both the R&D capital stock and unit cost in each industry.

In this paper, I present additional evidence on the nature and extent of R&D spillovers and on the connection between spillovers, R&D spending, and technological performance. I employ data from a survey of 350 R&D executives on the nature of appropriability and technological opportunity in 130 industries defined at the FTC Line of Business level. A more comprehensive account of the survey's findings concerning R&D appropriability may be found in Levin et al. (1987).

### III. New Evidence

Survey respondents were asked to rate, on a seven-point Likert scale, the effectiveness of seven methods of acquiring technical knowledge of process and product innovations developed by a competitor. Table 1 reports sample means for each question. In contrast to the set of questions studied by Cain Cockburn and Zvi Griliches (1988), nearly all questions in this group exhibit statistically significant industry effects in analyses of variance among industries defined at both the NSF (2.5 digit) and the FTC (3.5 digit) levels of aggregation.

The list of alternative methods of acquiring information about competitive technology, developed after extensive interaction with advisers and pretest subjects from the industrial R&D community, itself reveals

TABLE 1—EFFECTIVENESS OF ALTERNATIVE METHODS OF LEARNING

Method of Learning	Sample Means	
	Processes	Products
Licensing Technology	4.58 <sup>a</sup> (0.07)	4.62 <sup>a</sup> (0.07)
Patent Disclosures	3.88 <sup>a</sup> (0.05)	4.01 <sup>a</sup> (0.06)
Publications or Technical Meetings	4.07 (0.05)	4.07 <sup>b</sup> (0.05)
Conversations with Employees of Innovating Firm	3.64 <sup>b</sup> (0.06)	3.64 <sup>b</sup> (0.06)
Hiring Employees of Innovating Firm	4.02 <sup>a</sup> (0.07)	4.08 <sup>a</sup> (0.07)
Reverse Engineering of Product	4.07 <sup>a</sup> (0.07)	4.83 <sup>b</sup> (0.06)
Independent R&D	4.76 (0.06)	5.00 <sup>a</sup> (0.05)

Notes: 1 = not at all effective; 7 = very effective. Standard errors are shown in parentheses.

<sup>a</sup>Interindustry differences in means significant at .01 level.

<sup>b</sup>Interindustry differences in means significant at .05 level.

something important about the transmission of knowledge. Most methods of learning require some commitment of resources; contrary to their treatment in Spence's model, spillovers in R&D are not necessarily free. Indeed, the methods of acquiring technical information that appear on average to be most effective—licensing the technology, reverse engineering the product, and undertaking independent R&D—are likely to be the most costly. Most of the other channels of spillover—acquiring technical details through patent disclosures, publications, technical meetings, and informal conversations with rival employees—are relatively inexpensive.

The survey responses reveal a pattern of correlation among spillover channels that rely on interpersonal communication (publications and technical meetings, informal conversations, and hiring away employees). The effectiveness of these channels, and of reverse engineering as well, is negatively correlated with the time required to imitate an innovation. On the other hand, the effectiveness of licensing as a means of acquiring knowledge is positively correlated with the reported time required for imitation and with

the reported effectiveness of patents in preventing duplication. These relationships suggest that high scores on questions about the effectiveness of interpersonal channels, reverse engineering, and (more ambiguously) patent disclosures reflect the presence of involuntary (uncontrollable) spillovers of the sort that concerned Arrow and were modeled by Spence. A perception that licensing is an effective method of learning, however, should not be interpreted as a reflection of such spillovers; more likely, it signals that patents provide a reasonable degree of appropriability.

To relate the effectiveness of the various learning mechanisms to industry *R&D* investment and to technological performance, two approaches were employed. First, using the survey data at the level of the individual respondent, answers to the fourteen questions were reduced to three principal components, which had reasonably clear and meaningful interpretations. Standardized principal component scores were calculated for each respondent, aggregated to the FTC Line of Business level, and used to explain line of business *R&D* intensity (a three-year average, 1974–76, as reported to the FTC) as well as rates of process and product innovation (as reported by survey respondents). The second approach used cluster analysis of industry mean responses to group lines of business with similar patterns of response. Dummy variables representing cluster membership were then used to explain *R&D* intensity and innovation rates.

In the first approach, a variety of factor-analytic methods were employed to reduce the data, but none yielded perceptibly better results than a simple extraction of principal components. The first principal component loads most heavily on interpersonal channels of spillover, but it also gives substantial weight to learning from patent disclosures. Reverse engineering gets a smaller, but still positive, weight. The first component may thus be reasonably interpreted as measuring the effectiveness of low-cost means of acquiring technological information from competitors, or, nearly equivalently, as measuring the extent of involuntary spillovers. The sec-

ond component loads most heavily on learning through licensing, and the third loads almost exclusively on independent *R&D*.

Standardized scores on the first three principal components contribute effectively nothing to the explanation of interindustry variance in *R&D* intensity ( $F = 0.30$ ,  $R^2 = 0.01$ ). The first principal component reflecting the importance of low-cost channels of spillover is, however, significantly related to reported rates of process and product innovation ( $t = 2.96$  in both regressions). Together, the three principal components explain 11 percent of the variance in process innovation rates and 14 percent of the variance in product innovation rates.

The second approach produced similar results. I confine attention to the questions concerning learning about new processes since these questions yielded a clustering of industries with better statistical properties and a clearer interpretation than the questions about new products. Three clusters were found, as reported in Levin et al. The first cluster consists disproportionately of chemical and materials processing industries. The 68 lines of business in this cluster typically rely upon licensing and independent *R&D* to learn about competitive technology. The second cluster includes, among others, most high-technology industries other than chemicals and pharmaceuticals: semiconductors, computers, communications equipment, aircraft, guided missiles, measuring and controlling devices, and optical instruments. The 43 lines of business in this second cluster report that interpersonal channels and reverse engineering are most effective. The 19 industries in the third cluster, drawn predominantly from the food processing and metal working sectors, find that none of the mechanisms of learning is particularly effective.

Cluster assignment is not a significant determinant of *R&D* intensity ( $F = 1.80$ ,  $R^2 = 0.03$ ), but it does have a significant impact on reported rates of process innovation ( $F = 10.57$ ,  $R^2 = 0.15$ ) and product innovation ( $F = 9.38$ ,  $R^2 = 0.14$ ). In particular, the industries in the second cluster, in which interpersonal channels and reverse engineering

are most effective, had average rates of innovation that were significantly greater than those of industries relying upon licensing and independent R&D ( $t = 4.02$  for processes and  $t = 3.83$  for products), and greater than those of industries in which all learning mechanisms are weak ( $t = 3.73$  for processes and  $t = 3.46$  for products).

#### IV. Conclusions

Spence's model predicts that spillovers discourage R&D investment but may be conducive to rapid technical progress. The results presented here, though only suggestive, give some support to the latter hypothesis, but none to the former. It may be useful to ask why the disincentive effect of spillovers is not discernible in these data.

Consider the industries that reported the highest level of spillovers, measured by the sum of mean scores on the three questions involving interpersonal channels. At the 2.5 digit level of aggregation used in the NSF's annual R&D survey, the top four industries were computers, communications equipment, electronic components, and aircraft. Spence's model suggests that the disincentive effect would impinge most strongly in these industries, yet they all rank high in R&D intensity.

A possible explanation is that these primarily electronics-based industries face particularly abundant technological opportunities that offset the disincentive effect of spillovers. The data, however, do not support this hypothesis. When a vector of survey-based variables representing technological opportunity is added to the R&D regressions, these variables are highly significant, but both the principal components and the cluster identity variables remain statistically insignificant. (The reported relationships between appropriability variables and technological performance, however, remain statistically significant when the opportunity variables are added to the innovation rate equations.)

Another, more speculative, possibility is that interindustry differences in the nature of technical advance may explain why, contrary

to the prediction of Spence's model, R&D investment is not discouraged by the high levels of spillover in electronics-based industries.

In Spence's model, own and rival R&D are perfect substitutes. This may be a reasonable characterization of industries with "discrete" technologies, in which innovations more or less stand alone as isolated discoveries. In such a regime, knowledge of firm A's innovation may lower the marginal productivity of firm B's R&D investment, by rendering further effort duplicative or by diverting it elsewhere in the opportunity set. Spillovers would thus be expected to discourage R&D in industries with discrete technologies, such as the chemical and drug industries prior to the revolution in genetic engineering.

By contrast, technical advance in the electronics industries has been much more "cumulative" than "discrete." This period's microelectronic device incorporates many features developed in previous periods, and the new technology it embodies is typically a foundation for next period's innovations. When innovations are "building blocks" in this sense, spillovers of a rival's R&D may raise the marginal product of own R&D. In such a regime, a high degree of spillovers may not only spur technical advance but also encourage R&D investment.

#### REFERENCES

- Arrow, Kenneth J., "Economic Welfare and the Allocation of Resources for Invention," in Richard R. Nelson, ed., *The Rate and Direction of Inventive Activity*, Princeton: Princeton University Press, 1962, 609-25.
- Bernstein, Jeffrey and Nadiri, M. Ishaq, "Research and Development and Intraindustry Spillovers: An Empirical Application of Dynamic Duality," NBER Working Paper No. 2002, August 1986.
- Cockburn, Iain and Griliches, Zvi, "Industry Effects and Appropriability Measures in the Stock Market's Valuation of R&D and Patents," *American Economic Review Proceedings*, May 1988, 78, 419-23.

- Jaffe, Adam, "Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits and Market Value," *American Economic Review*, December 1986, 76, 984-1001.
- Levin, Richard C. et al., "Appropriating the Returns from Industrial R&D," *Brookings Papers on Economic Activity*, 3:1987, forthcoming.
- \_\_\_\_\_ and Reiss, Peter C., "Cost-Reducing and Demand-Increasing R&D with Spillovers," working paper, Stanford University, November 1986.
- Mansfield, Edwin, "How Rapidly Does New Industrial Technology Leak Out?," *Journal of Industrial Economics*, December 1985, 34, 217-24.
- \_\_\_\_\_, Schwartz, M., and Wagner, S., "Imitation Costs and Patents: An Empirical Study," *Economic Journal*, December 1981, 91, 907-18.
- Spence, Michael, "Cost Reduction, Competition, and Industry Performance," *Econometrica*, January 1984, 52, 101-21.

# Interindustry *R&D* Spillovers, Rates of Return, and Production in High-Tech Industries

By JEFFREY I. BERNSTEIN AND M. ISHAQ NADIRI\*

Firms undertaking *R&D* investment are unable to completely appropriate all of the benefits from their *R&D* projects. J. Schumpeter (1950), J. Schmookler (1966), R. E. Evenson and Y. Kislev (1973), N. Rosenberg (1979), Z. Griliches (1979), and M. Spence (1984) have pointed out that the degree of appropriability can influence both the causes and consequences of *R&D* investment. The *R&D* investment by a firm reduces its own production cost and, as a result of spillovers, costs of other firms are also reduced. R. C. Levin and P. C. Reiss (1984) estimated that a 1 percent increase in the *R&D* spillover caused average cost to decline by .05 percent. A. Jaffe (1986) estimated that when spillovers increased by 1 percent profit increased by .3 percent. Our paper (forthcoming) focused on intra-industry spillovers of four manufacturing industries. Generally, average cost declined by .2 percent in response to a 1 percent growth in *R&D* spillovers.

A common feature of the previous empirical studies was the measurement of *R&D* spillovers as a single variable. This meant that individual firms or industrial sources of spillovers could not be distinguished as to their relative significance in influencing production cost and factor demand. The first purpose of this paper is to investigate the effects of interindustry *R&D* spillovers in five high-tech industries where each industry is treated as a separate spillover source. This treatment of *R&D* spillovers allows us to

estimate a matrix characterizing the sources and beneficiaries of each interindustry spillover.

The *R&D* spillovers create a dichotomy between private and social rates of return to *R&D* capital. The second purpose of this paper is therefore to compute both the social and private rates of return to *R&D* capital. The private return is measured as the variable cost reduction in an industry due to its own *R&D* capital expansion. The social rate of return to an industry's *R&D* capital consists of the private rate plus the interindustry marginal cost reductions due to the spillovers generated by the industry's *R&D* capital. Since each industry has the potential to cause distinct spillovers on each of the other industries, the spillover components of the social rates of return are decomposed by externality-receiving industries.

## I. Cost, Factor Demand, and Interindustry Spillover

Production cost and factor demands of an industry are influenced by *R&D* capital accumulated by other industries. The existence of these spillovers and their effects on production processes can be estimated by specifying a variable cost function for each industry,

$$\begin{aligned}
 (1) \quad \ln c_v/w_m = & \beta_0 + \beta_l \ln \omega_l + \beta_p \ln \omega_p \\
 & + \beta_y \ln y + \beta_i \ln K_r^i + \beta_{ip} \ln \omega_l \ln \omega_p \\
 & + \beta_{ly} \ln \omega_l \ln y + \beta_{li} \ln \omega_l \ln K_r^i \\
 & + \beta_{py} \ln \omega_p \ln y + \beta_{pi} \ln \omega_p \ln K_r^i \\
 & + \beta_{yi} \ln y \ln K_r^i + (\ln K_r^i + \beta_{ls} \ln \omega_l \\
 & + \beta_{ps} \ln \omega_p) \sum_{\substack{j=1 \\ j \neq i}}^N \beta_{ij} \ln K_r^j + u_c,
 \end{aligned}$$

\*Department of Economics, Carleton University, Ottawa, Canada and Department of Economics, New York University, New York, 10003, respectively. We thank Erwin Diewert, Zvi Griliches, Peter Mohnen, and Peter Reiss for comments and discussions on the research reported in this paper. Financial support was provided by NSF grant PRA-8108635. Technical support from the C.V. Starr Center for Applied Economics is also gratefully acknowledged.

where  $c_v$  is variable cost,  $w_m$  is the factor price of materials,  $\omega_l = w_l/w_m$ , where  $w_l$  is the wage rate,  $\omega_p = w_p/w_m$ , where  $w_p$  is the rental rate on physical capital,  $y$  is output,  $K_r^i$  is the industry's own R&D capital and  $K_r^j$  is another industry's R&D capital,  $N$  is the number of industries, and  $u_e$  is the error term.

The variable cost function is a truncated translog, since there are no own or squared second-order terms in the function.<sup>1</sup> Each industry's R&D capital generates a distinct influence on the variable cost and factor demands of every other industry. In this model there are  $N-1$  spillovers for each industry and the sum of these spillovers or the pool of knowledge is given by  $\sum_{j=1, j \neq i}^N \beta_{ij} \ln K_r^j$ . The pool of knowledge is determined by the estimation of the model.

The existence of the parameters  $\beta_{ls}$  and  $\beta_{ps}$  enables the spillover to affect the labor, physical capital, and materials cost shares. We can see this by using Shephard's Lemma,

$$(2) \quad s_l = \beta_l + \beta_{lp} \ln \omega_p + \beta_{ly} \ln y \\ + \beta_{li} \ln K_r^i + \beta_{ls} \sum_{\substack{j=1 \\ j \neq i}}^N \beta_{ij} \ln K_r^j + u_l,$$

$$(3) \quad s_p = \beta_p + \beta_{lp} \ln \omega_l + \beta_{py} \ln y \\ + \beta_{pi} \ln K_r^i + \beta_{ps} \sum_{\substack{j=1 \\ j \neq i}}^N \beta_{ij} \ln K_r^j + u_p,$$

where  $s_l = w_l v_l / c_v$  is the labor variable cost share,  $v_l$  is the labor demand,  $s_p = w_p v_p / c_v$  is the physical capital variable cost share,  $v_p$  is the physical capital demand, the material cost share is  $s_m = 1 - s_l - s_p$  by definition, and  $u_l$  and  $u_p$  are the error terms. The two parameters  $\beta_{ls}$  and  $\beta_{ps}$  create the nonlinearity in the model, and in conjunction with the other spillover parameters ( $\beta_{ij}$ ) determine

the factor biases associated with the R&D spillovers. The equilibrium characterized by equations (1), (2), and (3) is short run. The R&D capital is assumed to be a quasi-fixed factor because of the development costs which generate lags in the completion of R&D projects. Thus short-run cost is not minimized with respect to R&D capital.<sup>2</sup>

The variable cost or productivity effects associated with each of the R&D spillovers are

$$(4) \quad \partial \ln c_v / \partial \ln K_r^j \\ = \beta_{ij} (\ln K_r^i + \beta_{ls} \ln \omega_l + \beta_{ps} \ln \omega_p), \\ j \neq i, j = 1, \dots, 5$$

The productivity effects are not constant but depend on the R&D capital of the spillover receiving industry, along with the relative factor prices. The parameter  $\beta_{ij}$  defines the distinct effect that the R&D capital from industry  $j$  exerts on the  $i$ th receiving industry.

The parameters  $\beta_{ls}$ ,  $\beta_{ps}$  transform the industry-specific productivity effects of the spillovers into factor bias effects.

$$(5) \quad \partial s_k / \partial \ln K_r^j = \beta_{ks} \beta_{ij}; \\ k = l, p, m, i \neq j, j = 1, \dots, 5$$

where  $\beta_{ms} = -(\beta_{ls} + \beta_{ps})$ . If a variable factor cost share increases (decreases, or does not change), then the interindustry spillover is variable factor using (reducing, or neutral). The effect on each variable factor demand consists of the sum of the productivity and factor bias effects. It is possible for the variable factors to be complementary, substitutable or independent of each of the interindustry R&D spillovers.

<sup>1</sup>The translog variable function was estimated but we were unable to find parameter estimates which satisfied the regularity conditions for this function.

<sup>2</sup>Physical capital and certain types of labor could also be quasi fixed. However, the focus of the paper is on R&D capital. The model captures the relative inflexibility of R&D capital compared to other factors of production.



## II. The Effects of R&D Spillovers

Five high-technology industries were analyzed in this paper: chemical products (SIC 28), non-electrical machinery (SIC 35), electrical products (SIC 36), transportation equipment (SIC 37), and scientific instruments (SIC 38). The sample period was 1958 to 1981 and most of the data for these industries were obtained from published sources of the Bureau of Economic Analysis (BEA).<sup>3</sup> The data consisted of gross output ( $y$ ) in 1982 dollars; labor ( $v_l$ ) was measured as total man-hours. The wage rate ( $w_l$ ) was defined as the ratio of adjusted total payroll to total man-hours. The physical capital stock ( $v_p$ ) was measured as net capital stock. The R&D capital stock ( $K_r$ ) was constructed by perpetual inventory method with a depreciation rate of 10 percent. Adjustments were made to avoid double counting with respect to capital used in R&D activities. The after-tax rental rates or factor prices of physical and R&D capital were derived using the familiar user-cost formula taking account of tax rules applicable to each asset. Real materials was defined as gross output minus real value-added. The materials price was implicitly calculated as the ratio of nominal materials to real materials.

Equations (1), (2), and (3) were estimated for each of the five industries. The sample period for the estimation was 1958 to 1981. The estimator used was full information maximum likelihood and the convergence criteria was .001. Each industry's own R&D capital and each of the spillovers were lagged one period in the estimation since all R&D capital stocks were exogenous in the model.<sup>4</sup>

The estimation of the model for each industry proceeded in the following manner. First, each spillover source industry was included in the model individually, next, pairs of spillover source industries were included, then groups of three, and lastly, all four

spillover source industries were included in estimating the model. The criteria for accepting an industry as a spillover source were the satisfaction of the regularity conditions pertaining to the variable cost function and the statistical significance of the parameters. The estimation results were quite stable in the sense that when a specific spillover source industry generated by itself either a statistically insignificant effect, or caused the violation of regularity conditions, these problems also occurred when the specific industry was grouped with the other source-industries. Thus the empirical results were robust, both in an economic and statistical sense, as to the acceptance of a particular industry as a source of spillover for each receiving industry.<sup>5</sup>

The effects of the spillovers are presented in Table 1. For example, for the chemical products industry, the R&D capital spillover emanated from the scientific instruments industry. Variable cost declined as a result of the spillover over the sample period. A 1 percent increase in the spillover caused variable cost in this industry to decline by .21 percent in 1961 and by .09 percent in 1981. Both labor and material demands declined in response to the spillover suggesting that these two factors were partly substituted by the R&D capital spillover. However, the demand for physical capital increased in response to the spillover. Physical capital was a complement to the spillover. It was estimated that the spillover was labor and material reducing, while physical capital using.<sup>6</sup>

Our empirical results point to some general findings. First, variable cost for each industry was reduced by R&D capital spill-

<sup>5</sup>The variable cost function was assumed to be homogeneous of degree one in variable factor prices. This was a maintained hypothesis as relative variable factor prices affected variable cost. This assumption was reasonable to maintain because it implies that each variable factor demand depends on relative and not nominal prices.

<sup>6</sup>Besides looking at the product of  $\beta_{ij}$  with each of  $\beta_{ls}$ ,  $\beta_{ps}$  and  $\beta_{ms} = -(\beta_{ls} + \beta_{ps})$ , if the factor demand elasticity is positive or if negative smaller in absolute value than the cost elasticity, then the spillover is factor using, otherwise it is factor reducing.

<sup>3</sup>For a more detailed discussion of the data, a longer version of the paper is available from the authors.

<sup>4</sup>Details of the estimation results are available upon request.

TABLE 1—SPILLOVER EFFECTS ON VARIABLE COST AND FACTOR DEMAND

Receiving Industry	1961	1971	1981
Chemical Products <sup>a</sup>			
Var. Cost	-0.208	-0.133	-0.089
Labor	-0.825	-0.797	-1.088
Phys. Cap.	3.918	2.506	1.420
Materials	-0.382	-0.312	-0.269
Non-Electrical Machinery <sup>b</sup>			
Var. Cost	-0.029	-0.036	-0.058
Labor	0.059	0.056	0.047
Phys. Cap.	-1.740	-1.969	-0.689
Materials	0.013	0.000	-0.017
Electrical Products <sup>a</sup>			
Var. Cost	-0.109	-0.117	-0.119
Labor	-0.109	-0.117	-0.119
Phys. Cap.	-0.109	-0.117	-0.119
Materials	-0.109	-0.117	-0.119
Transportation Equipment <sup>c</sup>			
Var. Cost	-0.113	-0.106	-0.092
Labor	-0.360	-0.346	-0.363
Phys. Cap.	3.637	4.544	1.398
Materials	-0.204	-0.197	-0.187
Scientific Instruments <sup>b</sup>			
Var. Cost	-0.051	-0.059	-0.078
Labor	-0.025	-0.031	-0.045
Phys. Cap.	-1.240	-0.846	-0.391
Materials	-0.012	-0.022	-0.041

<sup>a</sup>Source Industry: Scientific Instruments.<sup>b</sup>Source Industry: Chemical Products, Electrical Products, Transportation Equipment.<sup>c</sup>Source Industry: Non-Electrical Machinery.

overs. Second, the spillover for each receiving industry emanated from a very narrow range of industries. There was only a single spillover source for three industries, while for the other two industries, although they were affected by the three industries, the spillover effects were equal across source-industries. Third, in four of the five industries, there were factor bias effects associated with the spillovers. The bias effects for labor and materials were always in the same direction and in the opposite direction to physical capital.

### III. Rates of Return *R&D* Capital

The private rate of return is defined by the real value of the variable cost reduction due to an increase in an industry's own *R&D*.

TABLE 2—PRIVATE RATES OF RETURN

Industry	Year	<i>R&amp;D</i> Capital	Physical Capital
Chem. Prods.	1961	0.194	0.067
	1971	0.132	0.082
	1981	0.133	0.135
Non-Elec. Mach.	1961	0.160	0.075
	1971	0.267	0.085
	1981	0.240	0.136
Electr. Prods.	1961	0.201	0.075
	1971	0.150	0.084
	1981	0.224	0.139
Transp. Equip.	1961	0.085	0.071
	1971	0.095	0.086
	1981	0.119	0.117
Scient. Instr.	1961	0.168	0.080
	1971	0.173	0.083
	1981	0.161	0.118

Namely,  $\rho^i = -(\partial c_v^i / \partial K_r^i) p_r^i$ ,  $i = 1, \dots, N$ , where  $\rho$  is the gross private rate of return and  $p_r$  is the price of *R&D* capital. Since the model is characterized by a short-run equilibrium, then the rate of return on *R&D* capital is endogenous and not assumed to be related to its rental rate. However, for physical capital, since it is part of short-run cost minimization, its gross rate of return is exogenous and given by the ratio of the rental rate to the purchase price for physical capital. In Table 2, we present the estimated private rates of return to *R&D* capital and physical capital. In four of the five industries, the gross private rate of return on *R&D* capital was, on average over the sample period, 1.5 to 2 times greater than the rate on physical capital. In addition, in these four industries the private rates of return on *R&D* were generally the same and ranged from .15 to .20. In the transportation equipment industry, the rates of return on the two types of capital were virtually equal.

The social rate of return to *R&D* differs from the private rate because of the existence of *R&D* spillovers. The social rate is defined inclusive of the spillover and it is equal to the private rate plus the sum of the marginal real interindustry variable cost reductions. Thus,  $\gamma^i = \rho^i - \sum_{j=1, j \neq i}^N (\partial c_v^j / \partial K_r^i) / p_r^i$ ,  $i = 1, \dots, N$ , where  $\gamma$  is the social rate of return to *R&D* capital.

In Table 3 we present the social rates of return and their decomposition. First, the

TABLE 3—SOCIAL RATES OF RETURN

Industry Source and Receiving Industries:	1961	1971	1981
Chemical Products			
Non-El. Machinery	.062	.058	.126
Scient. Instr.	.025	.020	.032
Social R of R	0.281	0.210	0.291
Non-El. Machinery			
Trans. Equip.	.418	.316	.210
Social R of R	0.577	0.583	0.450
Electrical Products			
Non-El. Machinery	.024	.024	.062
Scient. Instr.	.010	.008	.016
Social R of R	0.235	0.182	0.302
Transp. Equipment			
Non-El. Machinery	.014	.013	.035
Scient. Instr.	.006	.004	.009
Social R of R	0.105	0.112	0.163
Scient. Instruments			
Chemical Products	.926	.505	.657
Electr. Products	.522	.429	.471
Social R of R	1.615	1.107	1.289

R&D capital from chemical products generates spillovers on non-electrical machinery and scientific instruments. From the third row in Table 3, we note that the social rate of return to R&D capital in the chemical products industry was .29 in 1981. Indeed, over the sample period, the social rate was 1.5 to 2 times the private rate of return to R&D capital. The non-electrical machinery industry only generated a spillover on the transportation equipment industry. However, the effect was quite large so that the social rate of return on R&D capital in the non-electrical machinery industry was 2 to 3 times the private rate of return. The R&D capital from electrical products and transportation equipment each affected the same industries as the chemical products industries. In both cases, the social rate was greater than the private rate, but only by about 10 to 20 percent. The last industry is scientific instruments, which affected the chemical products and electrical products industries. The R&D capital from scientific instruments generated substantial spillover effects on these two industries. In fact, the social rate of return was around 10 times the private rate.

The results on the social rates of return point out that each industry was a R&D capital spillover source. Four of the five industries generated spillovers on two industries. Our results on the differences between social and private rates of return were consistent with the findings of E. Mansfield et al. (1971), Jaffe, and our cited paper. However, the results show clearly that it is important to distinguish among industries as sources of R&D spillovers.

#### IV. Conclusion

In this paper we have estimated a spillover network linking the receiving and sending industries. The findings suggest that there were significant differences among industries as both spillover senders and receivers. Several issues would be of interest for further research. First, we would like to investigate the spillover linkages among other 2-digit manufacturing industries. Another issue of importance is to analyze the relationship between an industry's own R&D capital and the spillovers due to the R&D activities pursued by other industries. Lastly, the existence of significant intra- and interindustry spillovers has important implications both with respect to tax effects on R&D investment and for competition policy relating to joint R&D ventures. These issues have yet to be investigated.

#### REFERENCES

- Bernstein, J. I. and Nadiri, M. I., "Research and Development, and Intraindustry Spillovers: An Empirical Application of Dynamic Duality," *Review of Economic Studies*, forthcoming.
- Evenson, R. E. and Kislev, Y., "Research and Productivity in Wheat and Maize," *Journal of Political Economy*, November/December 1973, 81, 1309-29.
- Griliches, Z., "Issues in Assessing the Contribution of Research and Development to Productivity Growth," *Bell Journal of Economics*, Spring 1979, 10, 92-116.
- Jaffe, A., "Technological Opportunity and Spillovers of R&D," *American Economic Review*, December 1986, 76, 984-1001.

- Levin, R. C. and Reiss, P. C., "Tests of a Schumpeterian Model of R&D and Market Structure," in Z. Griliches, ed., *R&D, Patents and Productivity*, NBER, Chicago: University of Chicago Press, 1984.
- Mansfield, E. et al., *Research and Innovation in the Modern Corporation*, New York: W. W. Norton, 1971.
- Rosenberg, N., "Science, Invention and Economic Growth," *Economic Journal*, March 1974, 84, 90-108.
- Schmookler, J., *Invention and Economic Growth*, Cambridge: Harvard University Press, 1966.
- Schumpeter, J., *Capitalism, Socialism and Democracy*, 3rd ed., New York: Harper and Row, 1950.
- Spence, M., "Cost Reduction, Competition and Industry Performance," *Econometrica*, January 1984, 52, 101-21.

# IS IT MONEY OR CREDIT, OR BOTH, OR NEITHER?†

## Credit, Money, and Aggregate Demand

By BEN S. BERNANKE AND ALAN S. BLINDER\*

Most standard models of aggregate demand, such as the textbook IS/LM model, treat bank assets and bank liabilities asymmetrically. Money, the bank liability, is given a special role in the determination of aggregate demand. In contrast, bank loans are lumped together with other debt instruments in a “bond market,” which is then conveniently suppressed by Walras’ Law.

Much recent research provides reasons to question this imbalance. A growing theoretical literature, based on models with asymmetric information, stresses the importance of intermediaries in the provision of credit and the special nature of bank loans. Empirically, the instability of econometric money-demand equations has been accompanied by renewed interest in the credit-GNP relationship (see especially the work of Benjamin Friedman).

We have developed several models of aggregate demand which allow roles for both money and “credit” (bank loans). We present a particularly simple one, a variant of the textbook IS/LM model, in this paper.

Though it has a simple graphical representation like IS/LM, this model permits us to pose a richer array of questions than does the traditional money-only framework.

### I. The Model

The LM curve is a portfolio-balance condition for a two-asset world: asset holders choose between money and bonds. Tacitly, loans and other forms of customer-market

credit are viewed as perfect substitutes for auction-market credit (“bonds”), and financial markets clear only by price. Models with a distinct role for credit arise when either of these assumptions is abandoned.

Following James Tobin (1970) and Karl Brunner and Allan Meltzer (1972), we choose to abandon the perfect substitutability assumption and ignore credit rationing.<sup>1</sup> Our model has three assets: money, bonds, and loans. Only the loan market needs explanation. We assume that both borrowers and lenders choose between bonds and loans according to the interest rates on the two credit instruments. If  $\rho$  is the interest rate on loans and  $i$  is the interest rate on bonds, then loan demand is:  $L^d = L(\rho, i, y)$ . The dependence on GNP ( $y$ ) captures the transactions demand for credit, which might arise, for example, from working capital or liquidity considerations.

To understand the genesis of loan supply, consider a simplified bank balance sheet (which ignores net worth) with assets: reserves,  $R$ ; bonds,  $B^b$ ; loans,  $L^s$ ; and liabilities: deposits,  $D$ . Since reserves consist of required reserves,  $\tau D$ , plus excess reserves,  $E$ , the banks’ adding-up constraint is:  $B^b + L^s + E = D(1 - \tau)$ . Assuming that desired portfolio proportions depend on rates of return on the available assets (zero for excess reserves), we have  $L^s = \lambda(\rho, i)D(1 - \tau)$ , with similar equations for the shares of  $B^b$  and  $E$ . Thus the condition for clearing the loan market is

$$(1) \quad L(\rho, i, y) = \lambda(\rho, i)D(1 - \tau).$$

†Discussants: Charles Freedman, Bank of Canada; Charles I. Plosser, University of Rochester; Robert H. Rasche, Michigan State University.

\*Princeton University, Princeton, NJ 08544. We are grateful to the NSF for supporting this research.

<sup>1</sup>Blinder (1987) offers a model in which there is rationing and no substitute for bank credit.

The money market is described by a conventional LM curve. Suppose banks hold excess reserves equal to  $\varepsilon(i)D(1-\tau)$ .<sup>2</sup> Then the supply of deposits (we ignore cash) is equal to bank reserves,  $R$ , times the money multiplier,  $m(i) = [\varepsilon(i)(1-\tau) + \tau]^{-1}$ . The demand for deposits arises from the transactions motive and depends on the interest rate, income, and total wealth, which is constant and therefore suppressed:  $D(i, y)$ . Equating the two gives

$$(2) \quad D(i, y) = m(i)R.$$

Implicitly,  $D(i, y)$  and  $L(\rho, i, y)$  define the nonbank public's demand function for bonds since money demand plus bond demand minus loan demand must equal total financial wealth.

The remaining market is the goods market, which we summarize in a conventional IS curve, written generically as<sup>3</sup>

$$(3) \quad y = Y(i, \rho).$$

## II. Graphical Representation

Use (2) to replace  $D(1-\tau)$  on the right-hand side of (1) by  $(1-\tau)m(i)R$ . Then (1) can be solved for  $\rho$  as a function of  $i$ ,  $y$ , and  $R$ :<sup>4</sup>

$$(4) \quad \rho = \phi(i, y, R).$$

Finally, substitute (4) into (3) to get

$$(5) \quad y = Y(i, \phi(i, y, R)),$$

which, in deference to Don Patinkin (1956),

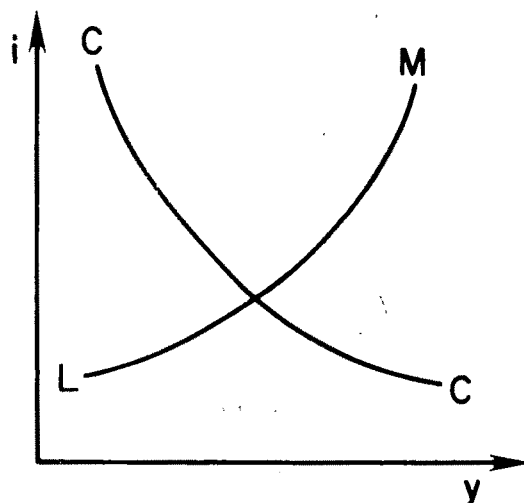


FIGURE 1

we call the CC curve (for “commodities and credit”). It is easy to see that the CC curve is negatively sloped like an IS curve, and for much the same reasons. However, it is shifted by monetary policy ( $R$ ) and by credit-market shocks that affect either the  $L(\cdot)$  or  $\lambda(\cdot)$  functions, while the IS curve is not. The CC and LM curves are shown together in Figure 1.

Our CC curve reduces to the IS curve if loans and bonds are assumed to be perfect substitutes either to borrowers ( $L\rho \rightarrow -\infty$ ) or to lenders ( $\lambda\rho \rightarrow \infty$ ), or if commodity demand is insensitive to the loan rate ( $Y\rho = 0$ )—which would make the loan market irrelevant to IS/LM. This clarifies the special assumptions implicit in the money-only view.

The opposite extreme, or credit-only view, would arise if money and bonds were perfect substitutes ( $D_i \rightarrow -\infty$ ), which would make the LM curve horizontal. Keynes' explanation for the liquidity trap is, of course, well known. We think of high substitutability as more likely to arise from financial innovations which create new money substitutes. However, even with a liquidity trap, monetary policy still matters because it influences the CC curve.

Now let us turn to the intermediate cases represented by Figure 1.

<sup>2</sup>For simplicity we assume that only  $i$ , not  $\rho$ , influences the demand for excess reserves.

<sup>3</sup>The interest rates in (3) should be real rates. But a model of aggregate demand takes both the price level and inflation as given; so we take the expected inflation rate to be constant and suppress it.

<sup>4</sup> $\rho$  is an increasing function of  $i$  as long as the interest elasticity of the money multiplier is not too large.

### III. Comparative Statics<sup>5</sup>

Most conventional shocks work in our model just as they do in IS/LM. For example, an expenditure shock shifts the CC curve along a fixed LM curve, and a money-demand shock shifts the LM curve along a fixed CC curve. The effects are familiar and need not be discussed. The only noteworthy difference is that a rise in bank reserves might conceivably raise the rate of interest in the credit model. Graphically, the ambiguity arises because an increase in  $R$  shifts both the CC and LM curves outward. Economically, the credit channel makes monetary policy more expansionary than in IS/LM and therefore raises the transactions demand for money by more than in the conventional model.

Greater interest attaches to issues that elude the IS/LM model. An upward shift in the credit supply function,  $\lambda(\cdot)$  (which might correspond, for example, to a decrease in the perceived riskiness of loans) shifts the CC curve outward along a fixed LM curve, thereby raising  $i$  and  $y$ . The interest rate on loans,  $\rho$ , falls, however. An upward shift in the credit demand function,  $L(\cdot)$ , which might correspond to a greater need for working capital, has precisely the opposite effects.

We find it difficult to think of or identify major shocks to credit demand, that is, sharp increases or decreases in the demand for loans at given interest rates and GNP. But shocks to credit supply are easy to conceptualize and to find in actual history. For example, Bernanke's (1983) explanation for the length of the Great Depression can be thought of as a downward shock to credit supply stemming from the increased riskiness of loans and banks' concern for liquidity in the face of possible runs. According to

the model, such a shock should reduce credit, GNP, and the interest rate on government bonds while raising the interest rate on loans. Another notable example with the same predicted effects is the credit controls of March-July 1980. In this instance "tight money" should, and apparently did, reduce interest rates on government bonds.

### IV. Implications for Monetary Policy

We turn next to the traditional target and indicator issues of monetary policy. The so-called monetary indicator problem arises if the central bank sees its impact on aggregate demand only with a lag but sees its impacts on financial-sector variables like interest rates, money, and credit more promptly. What does our model say about the suitability of money or credit as indicators?

Table 1 shows the qualitative responses of GNP, money, credit, and bond interest rates to a wide variety of shocks, assuming that bank reserves is the policy instrument. Columns 1 and 2 display a conclusion familiar from IS/LM: money is a good qualitative indicator of future GNP movements except when money demand shocks are empirically important. Columns 1 and 3 offer the corresponding conclusion for credit: credit is a good qualitative indicator except when there are important shocks to credit demand. If money demand shocks were indeed more important than credit demand shocks in the 1980's, credit would have been a better indicator than money.

What about the target question, that is, about the choice between stabilizing money vs. stabilizing credit? Rather than try to conduct a complete Poole-style (1970) analysis, we simply ask whether policymakers would respond "correctly" (i.e., in a stabilizing way) to various shocks if they were targeting money or targeting credit.

Consider first an expansionary IS shock. Table 1 (line 5) shows that both money and credit would rise if bank reserves were unchanged. Hence a central bank trying to stabilize either money or credit would contract bank reserves, which is the correct stabilizing response. Either policy works, at least qualitatively. A similar analysis applies

<sup>5</sup>Most comparative statics results require no assumptions other than the ones we have already made. But, in a few cases, we encounter theoretical ambiguities that can be resolved by invoking certain elasticity assumptions spelled out in a longer version of this paper. If output is fixed on the supply side,  $y$  would be replaced by  $P$  in Figure 1 and in the text discussion that follows.

TABLE 1—EFFECTS OF SHOCKS ON  
OBSERVABLE VARIABLES

Rise in:	(1) Income	(2) Money	(3) Credit	(4) Interest Rate <sup>a</sup>
Bank Reserves	+	+	+	—
Money Demand	—	+	—	+
Credit Supply	+	+	+	+
Credit Demand	—	—	+	—
Commodity Demand	+	+	+	+

<sup>a</sup> On bonds.

to shocks to the supply of credit or to the money multiplier.

But suppose the demand for money increases (line 2), which sends a contractionary impulse to GNP. Since this shock raises *M*, a monetarist central bank would contract reserves in an effort to stabilize money, which would destabilize GNP. This, of course, is the familiar Achilles heel of monetarism. Notice, however, that this same shock would make credit contract. So a central bank trying to stabilize credit would expand reserves. In this case, a credit-based policy is superior to a money-based policy.

The opposite is true, however, when there are credit-demand shocks. Line 4 tells us that a contractionary (for GNP) credit-demand shock lowers the money supply but raises credit. Hence a monetarist central bank would turn expansionary, as it should, while a creditist central bank would turn contractionary, which it should not.

We therefore reach a conclusion similar to that reached in discussing indicators: If money-demand shocks are more important than credit-demand shocks, then a policy of targeting credit is probably better than a policy of targeting money.

### V. Empirical Evidence

The foregoing discussion suggests that the case for credit turns on whether credit demand is, or is becoming, relatively more stable than money demand. We conclude with some evidence that this is true, at least since 1979.<sup>6</sup>

<sup>6</sup>In what follows, "money" is *M1*, "credit" is an aggregate invented by one of us: the sum of intermediated borrowing by households and businesses (derived

TABLE 2—SIMPLE CORRELATIONS OF GROWTH RATES  
OF GNP WITH GROWTH RATES OF  
FINANCIAL AGGREGATES, 1973–85<sup>a,b</sup>

Period	With Money	With Credit
1953:1–1973:4	.51,.37	.17,.11
1974:1–1979:3	.50,.54	.50,.51
1979:4–1985:4	.11,.34	.38,.47

<sup>a</sup>Growth rates are first differences of natural logarithms.

<sup>b</sup>Correlations in nominal terms come first; correlations in real terms come second.

Table 2 shows the simple correlations between GNP growth and growth of the two financial aggregates during three periods. Money was obviously much more highly correlated with income than was credit during the period of stable money demand, 1953–73. But the two financial aggregates were on a more equal footing during 1974:1–1979:3. Further changes came during the period of unstable money demand, 1979:4–1985:4; money-GNP correlations dropped sharply while money-credit correlations fell only slightly, giving a clear edge to credit.<sup>7</sup>

More direct evidence on the relative magnitudes of money-demand and credit-demand shocks was obtained by comparing the residuals from estimated structural money-demand and credit-demand functions like  $D(\cdot)$  and  $L(\cdot)$  in our model. We used the logarithmic partial adjustment model, with adjustment in nominal terms, which we are not eager to defend but which was designed to fit money demand. Hence, our procedure seems clearly biased toward finding relatively larger credit shocks than money shocks.

Unsurprisingly, estimates for the entire 1953–85 period rejected parameter stability across a 1973:4–1974:1 break, so we concentrated on the latter period.<sup>8</sup> Much to our

from Flow-of-Funds data). For details and analysis of the latter, see Blinder (1985).

<sup>7</sup>Similar findings emerged when we controlled for many variables via a vector-autoregression and looked at correlations between VAR residuals.

<sup>8</sup>Estimation was by instrumental variables. Instruments were current, once, and twice lagged logs of real government purchases, real exports, bank reserves, and a supply shock variable which is a weighted average of the relative prices of energy and agricultural products.



nagement, we estimated moderately sensitive money and credit demand equations for the 1974:1–1985:4 period on the first try (standard errors are in parentheses):

$$\log M = -.06 + .939 \log M_{-1} - .222i \\ (.34) \quad (.059) \quad (.089) \\ + .083 \log P + .012 \log y \\ (.052) \quad (.059)$$

$$SEE = .00811 \quad DW = 2.04,$$

$$\log C = -1.75 + .885 \log C_{-1} - .424p \\ (.063) \quad (.076) \quad (.285) \\ + .514i + .075 \log P + .292 \log y \\ (.389) \quad (.086) \quad (.107)$$

$$SEE = .00797, \quad DW = 2.44.$$

where  $y$  is real GNP,  $P$  is the GNP deflator,  $p$  is the bank prime rate, and  $i$  is the three-month Treasury bill rate. Although the interest rate coefficients in the credit equation are individually insignificant, they are jointly significant, have the correct signs, and are almost equal in absolute value—suggesting a specification in which the spread between  $p$  and  $i$  determines credit demand. Notice that the residual variances in the two equations are about equal.

Since the sample was too short to test reliably for parameter stability, we examined the residuals from the two equations over two subperiods with these results:

period	variance of money residual	variance of credit residual
1974:1–1979:3	$.265 \times 10^{-4}$	$.687 \times 10^{-4}$
1979:4–1985:4	$.888 \times 10^{-4}$	$.435 \times 10^{-4}$

The differences are striking. By this crude

measure, the variance of money-demand shocks was much smaller than that of credit-demand shocks during the first subperiod but much larger during the second.

The evidence thus supports the idea that money-demand shocks became much more important relative to credit-demand shocks in the 1980's. But that does not mean we should start ignoring money and focusing on credit. After all, it is perfectly conceivable that the relative sizes of money-demand and credit-demand shocks will revert once again to what they were earlier. Rather, the message of this paper is that a more symmetric treatment of money and credit is feasible and appears warranted.

## REFERENCES

- Bernanke, Ben S., "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review*, June 1983, 73, 257–76.
- Blinder, Alan S., "Credit Rationing and Effective Supply Failures," *Economic Journal*, June 1987, 97, 327–52.
- , "The Stylized Facts About Credit Aggregates," mimeo., Princeton University, June 1985.
- Brunner, Karl and Meltzer, Alan H., "Money, Debt, and Economic Activity," *Journal of Political Economy*, September/October 1972, 80, 951–77.
- Patinkin, Don, *Money, Interest, and Prices*, New York: Harper and Row, 1956.
- Poole, William, "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quarterly Journal of Economics*, May 1970, 2, 197–216.
- Tobin, James, "A General Equilibrium Approach to Monetary Theory," *Journal of Money, Credit and Banking*, November 1970, 2, 461–72.

# Monetary Policy Without Quantity Variables

By BENJAMIN M. FRIEDMAN\*

The collapse of the money-income relationship in the 1980's has thrown into question long-standing presumptions about the appropriate conduct of monetary policy. Before the 1980's, economists and policymakers had long debated the role that aggregate measures of money (or credit) should play in the monetary policy process. Although issues of a nonempirical nature were also important in this regard (for example, the desire for a system under which policymakers could be readily monitored and held accountable), the central issue was always the stability and reliability of the money-income relationship. Those who believed that it was highly stable typically sought to tie monetary policy more rigidly to fixed money growth targets, while those who doubted this stability sought to base monetary policy not just on money but on other variables too (credit, for example), and in any case to make the connection between policy actions and either money or any other specific variables more flexible.

What was at issue throughout this period, however, was mostly the short-run conduct of monetary policy, and therefore the short-run stability of the money-income relationship: fluctuations from quarter to quarter, or perhaps even year to year. Few economists or policymakers expressed doubts that the money-income relationship was highly stable over a time horizon as long as the average business cycle, and therefore few argued that money growth should not follow a narrowly specified trend over several years taken together. For those who were skeptical that a more activist policy could successfully carry out countercyclical stabilization anyway, the widely agreed upon stability of the money-

income relationship over longer horizons led naturally to a fixed money growth policy even in the short run.

The events of the 1980's have been so important for thinking about monetary policy precisely because they have contradicted this more fundamental confidence in the stability of the money-income relationship in the longer run. For the five years ending a mid-1987, the average growth rate of the *M* money stock was 10.8 percent per annum—far above that for any sustained period in recent U.S. experience. Yet inflation has been modest by historical standards, and real income growth for this period as a whole has hardly been extraordinary compared to previous U.S. business cycle expansions. It is difficult to escape the conclusion that, not just for a year or a calendar quarter but over an entire half-decade, money growth has simply been irrelevant to any outcome that matters for monetary policy.

Analogous relationships between income or prices and other financial quantity variables have fared little or no better during this period. Broader measures of money, or the monetary base, or measures of credit have all fluctuated in patterns bearing little visible connection to any plausible objective of monetary policy. As a result, the entire role of such quantity variables in the monetary policy process—either money or any of the others—is now practically devoid of empirical support based on recent experience. At the same time, however, no one has satisfactorily outlined an alternative monetary policy framework that does not rely on such variables. The result is a vacuum at the center of the monetary policy process.

## I. Money and Income, Money and Prices

One picture and one example from the recent literature are sufficient to place in perspective the collapse of the relationship between money and either income or price in the 1980's.

\*Harvard University, Cambridge, MA 02138. I am grateful to Kenneth Kuttner and James Stock for helpful discussions, and to the Harvard Program for Financial Research and the NSF for research support.

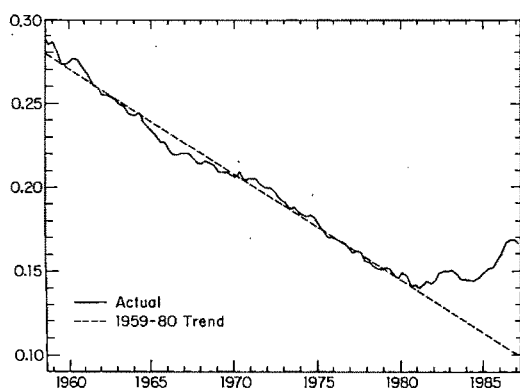


FIGURE 1. RATIO OF MONEY TO GNP

First, the picture: Figure 1 plots the ratio of the  $M1$  money stock to nominal GNP for each quarter from 1959:I (when the redefined  $M1$  series begins) to 1987:II. Through 1980:IV, the money-income ratio followed the familiar 3 percent per annum downward trend that practical discussions of monetary policy had come to treat as if it were a natural constant, with a standard deviation around the trend of only .0044 compared to a 1980:IV value of .1466. Since 1980, the short-run fluctuations have been visibly wider. More importantly, the downward trend has not just disappeared but reversed course. A simple extrapolation of the 1959–80 trend implies a money-income ratio of .1007 by 1987:II. The actual value was .1686, greater by more than 15 standard deviations.

The analogous relationship for credit, the outstanding indebtedness of all domestic nonfinancial borrowers, has fallen apart just as badly. During 1959–80, the credit-income ratio exhibited a standard deviation of only .0187 (around a negligible and statistically insignificant trend) compared to a 1980:IV value of 1.3782. By 1987:II the gap between the actual ratio and the trend extrapolation was more than 23 standard deviations.

Second, the example: In these same *Papers and Proceedings* four years ago, Milton Friedman (1984) argued that neither the money-income nor the money-price relationship had broken down after October 1979, when the Federal Reserve experimented with a policy centered on money growth targets. He argued instead that both relationships

had continued to hold up if interpreted correctly. For the money-income relationship, Friedman emphasized short-run comovements, focusing on each quarter's income growth and money growth in the prior quarter. For the money-price relationship, he emphasized longer-run comovements, focusing on average inflation over successive two-year intervals and average money growth over the prior two years.

Since Friedman wrote, however, both of the relationships on which he based his arguments have even changed sign. The correlation between the respective growth rates of nominal income and  $M1$  lagged one quarter was .45 during the 1979:IV–1983:IV sample he used. The same correlation computed for 1984:I–1987:II is *minus* .10. Friedman did not report a correlation for the biennial growth rates of prices and lagged money, but simply showed the data for each successive biennium, beginning with 1973:III–1975:III for the GNP deflator and 1971:III–1973:III for  $M1$ . The direction of the change in  $M1$  growth in each of these periods had foretold the direction of the change in inflation in the next, and on this basis Friedman predicted, “The increased rate of money growth in the 1981–83 biennium suggests that we have passed the trough in inflation and that inflation will be decidedly higher from 1983 to 1985 than it was from 1981–83” (p. 400). Instead, inflation turned out to be lower during 1983:III–1985:III than during 1981:III–1983:III, and it was lower still during 1985:III–1987:III despite continued high money growth during 1983:III–1985:III. The correlation computed over the five observations Friedman exhibited was .70. Computed over those five observations and the two more that are now available, the correlation is *minus* .23.

## II. Money and Credit as Information Variables

The breakdown of such simple money-income or money-price relationships casts doubt on the use of money (or credit) as a target of monetary policy in any rigid, mechanical sense. It need not preclude a useful role for such variables in the monetary policy process, however, as long as their movements provide information about subsequent

fluctuations of income or prices, or any other outcomes that monetary policy seeks to affect.<sup>1</sup> A policy framework based on aggregate measures of money (or credit) used as "information variables" is more flexible, and hence more complicated and harder to monitor externally, than a framework based on such variables used as policy targets. The greater the extent to which the relationships that connect these variables to income and prices are affected both by other variables (like interest rates) and by stochastic shocks, however, the greater are the relative merits of an information variable approach compared to a simpler targeting approach.

The events of the 1980's have even undermined the empirical foundation for basing monetary policy on financial quantity variables in this more flexible sense. Kenneth Kuttner and I (1988) have shown that evidence of a variety of forms, connecting money (or credit) to income and prices, has progressively deteriorated since 1979.

Table 1 shows  $\bar{R}^2$  statistics for the estimation of "St. Louis" equations relating the quarterly growth rate of nominal income to lagged growth rates of several respective financial quantity variables and the lagged growth rate of high-employment federal spending, over three sample periods.<sup>2</sup> For 1960:II–1979:III (i.e., until the introduction of the new monetary policy procedures), these equations all exhibit the familiar modest success in accounting for quarterly income growth, with  $\bar{R}^2$  values ranging from a low of .23 for the monetary base to a high of .32 for  $M1$ . Extending the sample to include data through year-end 1986 sharply lowers the  $\bar{R}^2$  in each case, however. Dropping the observations from the 1960's eliminates it almost altogether. Not one of these equations for the more recent period exhibits  $\bar{R}^2$  even as high as .10.

Table 2 shows  $F$ -statistics for tests of the null hypothesis that all of the coefficients on

TABLE 1—COEFFICIENT OF DETERMINATION FOR NOMINAL INCOME EQUATIONS

	1960:II– 1979:III	1960:II– 1986:IV	1970:III– 1986:IV
Monetary Base	.23	.15	.02
$M1$	.32	.11	.02
$M2$	.27	.19	.06
$M3$	.27	.16	.09
Credit	.28	.10	–.02

TABLE 2— $F$ -STATISTICS FOR INFORMATION VALUE OF MONEY ( $M1$ )

	1960:II– 1979:III	1960:II– 1987:II	1970:I– 1987:II
Fiscal Variable Excluded			
$Y$	6.16 <sup>a</sup>	2.63 <sup>b</sup>	1.42
$X$	1.98	1.91	1.33
$P$	3.62 <sup>b</sup>	.68	.47
Fiscal Variable Included			
$Y$	5.99 <sup>a</sup>	2.83 <sup>b</sup>	1.92
$X$	2.17 <sup>c</sup>	2.21 <sup>c</sup>	1.91
$P$	3.65 <sup>b</sup>	.75	.88

Note:  $Y$  = nominal GNP;  $X$  = real GNP;  $P$  = GNP price deflator.

<sup>a</sup>Significant at .01 level.

<sup>b</sup>Significant at .05 level.

<sup>c</sup>Significant at .10 level.

lagged  $M1$  growth are zero in equations from several series of vector autoregressions.<sup>3</sup> As in Table 1, results are shown for each of three sample periods: from the beginning of the  $M1$  series until the introduction of new monetary policy procedures, then through the most recent data available as of the time of writing, and then for the most recent data without the 1960's.

In the context of the information variable approach to monetary policy, the much debated issue of whether statistical experiments like these constitute valid tests of "causality" is beside the point. What matters is simply whether the movements of some financial quantity convey information about future movements of income or prices that is not already contained in observed movements of income or prices themselves. If so,

<sup>1</sup>See, for example, John Kareken et al. (1973) and my papers (1975, 1983).

<sup>2</sup>These equations differ from the St. Louis specification only by omitting the contemporaneous value of each independent variable.

<sup>3</sup>Each autoregression includes four lags on each variable in the system, plus a constant. All variables are in log differences.

then monetary policy can exploit that information by systematically reacting to observed movements of these variables, regardless of whether this information reflects true causation, reverse causation based on anticipations, or mutual causation by some independent but unobserved force.

As of 1979, the available evidence strongly supported the view that observed fluctuations of *M1* in the United States did contain such information about future movements of U.S. income and prices. By contrast, the same experiments carried out with data for the most recent 18 years provide no support for the view that fluctuations in *M1* carry information about future income and prices that is not already contained in fluctuations of income and prices themselves. Not one of the *F*-statistics for this more recent sample is significant at even the .10 level. Once again, what is true for *M1* is also true for other money and credit aggregates. The *F*-statistics for analogous experiments carried out with *M2* or credit in place of *M1* show the same pattern of changing significance as in Table 2. Not one of the *F*-statistics for *M2*, and not one for credit, is significant at the .10 level for the 1970:I–1987:II sample.

Not surprisingly, such findings have prompted a search for ways to “fix up” this form of test of the money-income relationship, just as a much more intensive search, which began even earlier, has sought to fix up the money demand function. James Stock and Mark Watson (forthcoming), for example, showed that with the right specification, lagged *M1* was in fact significant in equations for real income (proxied by industrial production) in tests based on monthly data for 1960:2–1985:12. For a system including money, income, prices, and an interest rate, together with a time trend, they reported an *F*-statistic of 3.04 (easily significant at the .01 level) for the null hypothesis that all of the lagged money coefficients were zero. As Kuttner and I have shown, however, merely extending the sample for this experiment through 1987:9 reduces the *F*-statistic to 1.80, just barely significant at the .10 level (*p* value .0994), and changes Stock and Watson’s results for the other systems that they investigated as well.

TABLE 3—DICKEY-FULLER *T*-STATISTICS FOR COINTEGRATION TESTS

	1959:I– 1979:III	1959:I– 1982:II	1959:I– 1987:III
Monetary Base	–2.90	–3.03	–0.22 <sup>a</sup>
<i>M1</i>	–1.53 <sup>a</sup>	–1.61 <sup>a</sup>	–0.34 <sup>a</sup>
<i>M2</i>	–3.67 <sup>b</sup>	–3.40 <sup>a,b</sup>	–2.69 <sup>a</sup>
Credit	–3.60 <sup>b</sup>	–3.28 <sup>c</sup>	–0.09 <sup>a</sup>

<sup>a</sup>Augmented Dickey-Fuller *t*-statistic.

<sup>b</sup>Significant at .05 level.

<sup>c</sup>Significant at .10 level.

Table 3 shows that the most recent experience has also eliminated statistical support for the hypothesis that income and money (or credit) are cointegrated. The table shows Dickey-Fuller *t*-statistics for the null hypothesis of no cointegration between nominal income and each of several financial quantity variables, in the presence of a possibly nonlinear time trend.<sup>4</sup> The results shown are based on quarterly data for three samples, which here differ only in their respective endpoints: before the introduction of new monetary policy procedures, before the abandonment of those procedures, and the latest data available as of the time of writing. At least for *M2* and credit, the data through 1979:III warranted rejecting the null hypothesis of no cointegration with nominal income at the .05 level. The data through 1982:II did so as well, albeit only at the .10 level for credit. For data through 1987:II, however, there is no evidence of cointegration with nominal income for any of these financial quantity variables.<sup>5</sup>

<sup>4</sup>The cointegrating equation is in each case  $\ln(f_t) = \ln(a + b \cdot t) + c \cdot \ln(y_t) + e_t$ , where  $f$  is the financial quantity,  $y$  is nominal income and  $e$  is a disturbance term. The null hypothesis of no cointegration means that  $e$  is nonstationary. The values shown are augmented Dickey-Fuller *t*-statistics in cases in which higher-order autocovariance of  $e$  is present, and ordinary Dickey-Fuller *t*-statistics otherwise.

<sup>5</sup>Tests carried out in the forms  $\ln(f_t) = a + b \cdot \ln(y_t) + e_t$  and  $(f_t/y_t) = a + b \cdot t + e_t$ , also show no evidence of cointegration for any of these financial quantity variables in the data through 1987:II.

### III. Questions About Monetary Policy Since 1982

If it is difficult to escape the conclusion that financial quantity variables have lost their relevance for monetary policy in the 1980's, it is also difficult to escape the conclusion that the Federal Reserve System has responded to this development by conducting monetary policy primarily with reference to short-term nominal interest rates (and, indirectly, dollar exchange rates). One reason for drawing this conclusion is simply the return to interest rate stability after the Federal Reserve "suspended" its *M1* target in 1982. The standard deviation of the month-to-month change in the three-month U.S. Treasury bill rate rose from .42 percent during 1970:1–1979:9 to 1.54 percent during 1979:10–1982:9, and then fell to .32 percent during 1982:10–1987:9. Another reason is that what movements in short-term interest rates have occurred since mid-1982 have shown little apparent connection to fluctuations of the major monetary aggregates (or credit) or to deviations of these aggregates from the corresponding official target ranges.

The success of U.S. monetary policy in macroeconomic terms during these years notwithstanding, a return to approximately the same monetary framework that the Federal Reserve employed a quarter-century ago should give cause for some concern—not least because of the systematic errors that the Federal Reserve made under that policy. The extensive analysis of U.S. monetary policy during the first two decades or so following the Treasury-Federal Reserve Accord, including research carried out at the time as well as subsequently, has documented three problems in particular. Each bears renewed consideration now that the Federal Reserve has returned to what amounts to a policy framework centered on controlling nominal interest rates.

First, and most obviously, this framework had no nominal quantity to anchor the price level. Although inflation was not therefore inevitable, there was little protection against it when inflationary pressures intensified in the late 1960's and especially in the 1970's. For some years following Thomas Sargent and Neil Wallace's (1975) demonstration that

basing monetary policy on nominal interest rates left the price level indeterminate in a model with "rational" expectations and perfectly flexible prices, many economists eschewed analysis of such a policy framework altogether, and concentrated only on policies based on controlling money. As Bennett McCallum (1981) has shown, however, even in Sargent and Wallace's model price indeterminacy results only when the central bank takes no account of prices (or any other nominal variable) in choosing the level at which to set interest rates.<sup>6</sup> Especially in a context that allows for rigidities in price setting behavior as well as more realistic representations of expectations, no one knows to what extent it is practically possible to avoid inflation with a monetary policy framework based on nominal interest rates or how best to structure such a policy to achieve that end.

Second, once inflation did emerge, Federal Reserve officials (and many other people too) often failed to distinguish nominal from real interest rates. As a result, they often associated higher observed interest rates with a tighter policy stance even when the increase in nominal interest rates merely kept pace with, or even fell short of, rising inflation expectations. In light of the enormous attention subsequently devoted to the distinction between nominal and real interest rates, both in the research literature and at the popular level, it would be surprising to see this mistake repeated in such an obvious way. Nevertheless, inferring "*the* real interest rate" is hardly straightforward. Expectations of future inflation are unobservable, and different people may hold different expectations anyway. Different people and different institutions also face different tax rates.

Third, there is also substantial evidence that, when U.S. monetary policy relied primarily on nominal interest rates in the past Federal Reserve officials systematically con-

<sup>6</sup>What McCallum actually showed was that taking account of *money* in setting the interest rate resolves the price indeterminacy. His result readily generalizes to the inclusion of any nominal variable, however.

fused the level of interest rates as the operating instrument of policy with the level of interest rates as an ultimate objective of policy.<sup>7</sup> As a result, they usually delayed too long before raising or lowering interest rate levels, and even then made changes of insufficient magnitude. Although this error too has received enormous attention, more in the research literature than in popular discussions, no one knows whether it is now possible to design a monetary policy framework based primarily on interest rates that can provide adequate safeguards against repeating it. Still less has anyone laid out in any detail what such safeguards might be.

The evidence from recent experience is clear on the potential role of financial quantity variables in the monetary policy process, and it is not positive. Perhaps the time has come for economists to turn at least some of the effort they are now spending on trying to overturn the evidence on these variables toward thinking about how best to conduct monetary policy without them.

<sup>7</sup>See, for example, Karl Brunner and Allan Meltzer (1964).

## REFERENCES

- Brunner, Karl and Meltzer, Allan H., *The Federal Reserve's Attachment to the Free Reserve Concept*, Washington: USGPO, 1964.
- Friedman, Benjamin M., "Targets, Instruments and Indicators of Monetary Policy," *Journal of Monetary Economics*, October 1975, 1, 443-73.
- , "The Roles of Money and Credit in Macroeconomic Analysis," in James Tobin, ed., *Macroeconomics, Prices and Quantities: Essays in Memory of Arthur M. Okun*, Washington: The Brookings Institution, 1983.
- and Kuttner, Kenneth N., "Money, Income and Prices after the 1980s," mimeo., NBER, 1988.
- Friedman, Milton, "Lessons from the 1979-82 Monetary Policy Experiment," *American Economic Review Proceedings*, May 1984, 74, 397-400.
- Kareken, John, H., Muench, Thomas and Wallace, Neil, "Optimal Open Market Strategy: The Use of Information Variables," *American Economic Review*, March 1973, 63, 156-72.
- McCallum, Bennett T., "Price Level Determinacy with an Interest Rate Policy Rule and Rational Expectations," *Journal of Monetary Economics*, November 1981, 8, 319-29.
- Sargent, Thomas J. and Wallace, Neil, "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *Journal of Political Economy*, April 1975, 83, 241-54.
- Stock, James H. and Watson, Mark W., "Integrating the Evidence on Money-Income Causality," *Journal of Econometrics*, forthcoming.

# Money and Credit in the Monetary Transmission Process

By KARL BRUNNER AND ALLAN H. MELTZER\*

The title of this session asks about the roles of money or credit in the transmission of monetary and real shocks. Our answer, repeated in different forms for more than two decades, is that the analysis of the transmission process is incomplete without both the money and credit markets and their interaction.

For many years, economists ignored the role of the credit markets. Recently, there has been some change. Concerns about financial fragility, banking failures, debt default and loan rationing focussed attention on credit markets. Reexamination of experience during the early 1930's (Ben Bernanke, 1983) raises an issue about whether credit market shocks operated 1) as an independent, or exogenous, impulse supplementing and reinforcing the monetary decline, or 2) as part of the interaction of credit, money (and other financial) markets.

Standard macroeconomic analyses, represented by the IS/LM system, cannot cope with these issues. The IS/LM system is restricted to a single portfolio equation representing all asset markets, so it cannot recognize the operation of an independent credit market. Consequently, the discussion of problems associated with the credit market typically proceeds outside the major macroeconomic paradigm. This can be seen in the work of Benjamin Friedman (1983), Joseph Stiglitz and Andrew Weiss (1981) and Bernanke. Their discussions and empirical work rely on *ad hoc* constructions and arguments not integrated into a broader macroeconomic scheme. The failure of integration reflects a judgment about the relevance of the dominant paradigm for analysis of the transmission of shocks from the credit market to the real economy and conversely.

This paper briefly considers three issues. First, we summarize our view of the interrelation of credit, money, and output markets. Then, we discuss some issues about banking and debt crises in the Great Depression discussed in Bernanke. Finally, we consider whether loan and equity rationing are a central feature of the transmission of monetary impulses, particularly deflationary impulses, as suggested by Stiglitz and Weiss, and by Stiglitz (1987).

## I. Money and Credit

The basic problem of the IS/LM framework follows from the implicit assumptions about asset substitution made to fit the world into a framework admitting only two assets (see Brunner, 1971; our papers, 1987; 1988). The analysis proceeds in one of two ways. Either money is a substitute only for financial assets ("bonds"), or there is general substitution over all assets. A two-asset world is achieved in the latter case by making financial and real assets perfect substitutes, or by restricting the analysis to episodes with comparatively small relative changes in the market conditions for financial and real assets. Consequently, the IS/LM framework is either empirically falsified or it fails to offer a useful explanation of major events. To study the interaction between money, financial assets, and real assets, we require a more inclusive analysis which explicitly incorporates a second asset market. The joint determination of bank credit, money stock, interest rate, and the price level of real assets may be achieved in this way.

The extended analysis supplements the money-market equation with an equation describing the credit market. The two-asset markets interact with the output market. Monetary impulses are conveyed by a process of general substitution driven by changes in relative prices 1) of various assets, 2) of

\*University of Rochester, Rochester, NY 14627 and Carnegie-Mellon University, Pittsburgh, PA 15213, respectively.



real assets and their services, and 3) of assets and output. This process extends the links between monetary impulses and the output market substantially beyond the narrow channel recognized in the IS/LM framework. In the extended model, impulses are conveyed simultaneously by a "Keynesian" channel (i.e., adjustments of interest rates on financial assets), and also by adjustments of the price level of real assets. The pattern of transmission depends, among other things, on the public's perception about the durability and persistence of various shocks. As shown most recently in our paper (1988), the analysis can be readily extended to include a wide range of intermediaries.

The interaction of the money market with a credit market modifies the results derived from the IS/LM model in several ways. First, IS/LM analysis of the transmission process necessarily emphasizes the magnitude of the interest elasticity of the demand for money. Interaction with the credit market changes this result. The *relative* magnitude of the interest elasticities on the two-asset markets, irrespective of their absolute value, determines the transmission of shocks by means of asset price adjustments. Second, the IS/LM model implies that an interest target policy effectively isolates output from shocks to money demand. This proposition is denied once we incorporate a credit market and admit credit market shocks. A strategy of monetary control achieves better results in response to credit market shocks than a strategy of interest control. Third, the effects of credit market shocks may differ in sign and magnitude from the effects produced by money market shocks. Fourth, the idea of a liquidity trap, suspending all connection from the money market to the output market, is firmly anchored in the structure of the IS/LM system. This idea is untenable once we include interaction with a credit market. Fifth, changes in reserve requirements if fully offset by open market operations impose no adjustments on asset markets and output market in an IS/LM analysis. The extended analysis shows that compensated changes in reserve requirements modify conditions on the credit market and thus induce adjust-

ments in the asset markets. Finally, the extended analysis offers a better framework for analyzing problems posed by regulation and deregulation.

Our summary reveals an important fact. In our analysis, the transmission of impulses (or shocks) to output depends on the operation and properties of the credit market. This holds both for monetary and real shocks, say, due to variations in the expected net return on real assets. An entirely separate, but related, issue is whether credit market shocks are an independent disturbance, as in Bernanke's discussion of the 1930's, or part of the process transmitting real and monetary shocks.

## II. Banking Crises

Bernanke reconsiders the role of banking crises in the propagation of depressions. Such crises are immediately reflected by a run on banks expressed by an increase in the ratio of currency to deposits. In the money-credit market analysis, the rise in the currency ratio lowers both the volume of bank credit and the money stock. Bank credit responds with greater sensitivity than the money stock. The reason is that the asset multiplier, linking the monetary base with bank credit, responds more sensitively to variations in the currency ratio than does the monetary multiplier (linking the base with the money stock). The effects on the asset markets are transmitted to the output market. Bernanke emphasizes correctly that these adjustments lower the degree of financial intermediation. Transaction and information costs of financial operations increase. The network of credit shrinks, and aggregate real demand for output and monetary velocity decline.

This account is incomplete. The run on banks and the resulting banking crisis would be avoided if the monetary authorities function as "lender of last resort." Their failure to do so raises the marginal productivity of the banks' reserve position, further reducing asset and monetary multipliers. Bank credit and money stock suffer a further reduction with corresponding repercussions on the output market. An unchecked run produces,

with some probability, bankruptcies and closures of banks. The probability of a run rises in the absence of a lender of last resort; the total demand for reserves exceeds the outstanding stock. Interest rates rise and asset prices fall, lowering asset values. The money stock and bank credit contract further. Initially, nonbank lenders may partially substitute, at rapidly rising marginal cost, for the decline in bank credit, but the net effect will be dominated by the decline in bank credit and disintermediation.

The financial crisis, revealed by bankruptcies and closures, has further consequences. It generates a large and pervasive uncertainty. This lowers the expected net real return on real assets. The decline in expected net real returns affects both asset and output markets. The adjustment imposed on asset markets and the interaction of asset and output markets reinforce the direct effect of bankruptcies and uncertainty on the output market.

This account of banking crises shows that their consequences depend on the working of a credit market and its interrelation with a money market. The credit market plays a major role in the conversion of the initial run on banks, via bankruptcies and bank closures, into a major deflationary process and, with the failure of the lender of last resort, into a possible banking crisis.

A question of interpretation remains. Are runs on banks the result of a cyclic decline or a consequence of monetary retardation? A comparison between the United Kingdom and the United States is informative. This comparison suggests that the observed differences in the two countries depend on the central bank's commitment to act as a lender of last resort and on the nature of the banking structure. An understanding by the banks and the public that the central bank accepts such a commitment moderates fears and uncertainties and avoids the subsequent banking crisis. In the United Kingdom after 1866, the central bank functioned as lender of last resort. There were no banking crises (see also Anna Schwartz, 1987). There was no central bank in the United States before 1914. The Federal Reserve refused in 1930–1933 to honor the commitment to serve as a

lender of last resort. Further, the repetitive occurrence of banking crises in the United States suggests that crises may be conditioned by the magnitude and virulence of the downswing.

We conclude that runs on banks and banking crises are endogenous events, conditional on the monetary propagation mechanism. The relevant conditions include the operation of a central bank, the structure of the banking system and the magnitude of the recession. Phillip Cagan's (1965) observation that banking crises typically occur late in the cycle and not at the beginning of the downswing offers some support for our conjecture. It follows under the circumstances that monetary policy, understood as a choice of institutions characterizing central banks and banking, shapes the likelihood and the pattern of potential banking crises.

Bernanke introduces the debt crisis as a separate and independent phenomenon, in addition to the banking crisis. He presents an impressive array of facts revealing the depth and pervasiveness of the debt crisis during the Great Depression. The transmission of the monetary retardation initiated in 1929, amplified by the banking crises appearing in the 1930's, lowered the price level of output and, even more, the price level of real assets. This massive deflation occurred in the context of an extensive network of private debt accumulated during the 1920's. The net worth position of households and business firms fell. Given the distribution of debtor positions, the deflationary process necessarily increased the number of bankruptcies, lowering the net worth of creditors and accelerating the debt crisis fostered by falling prices (and incomes). The risk premium on many assets rose, further reducing prices on real assets. (Interest rates on (default) risk-free securities declined due to an allocational shift from real assets to such securities.) These adjustments, unleashed on the asset markets, reinforced the direct effect of monetary contraction on the aggregate real demand for output.

Our discussion makes clear that we accept Bernanke's emphasis on the role of the debt crisis as an important component of the propagation mechanism. We do not accept

Bernanke's analysis of the debt crisis as a separate and independent exogenous shock. Once the monetary authorities allow the emergence of a major deflation of asset and output price levels, in a system with many holders of nominally fixed debt, a debt crisis is an induced response to the deflation. A minor debt crisis occurred in the United States early in the 1980's mainly as a result of a lower, positive rate of inflation.

This account, explicitly acknowledging the role of debt and credit in the propagation of major depressions, removes an objection to the monetary explanation of the Great Depression. The observation that real balances rose and velocity fell during the early 1930's is said to disconfirm the thesis of a (possible partial) monetary shock. The banking and debt crisis, unleashed in the propagation of such a shock through the economy under prevailing monetary arrangements, explains the emergence of the relatively large decline in velocity. This, in turn, explains why the deflation and decline are disproportionately large relative to the decline in the money stock. The secondary and tertiary effects of the monetary retardation, transmitted through the money-credit process and augmented by the failure of the lender of last resort, magnified the response to the monetary decline and induced an endogenous flight to money large enough to raise real cash balances. Despite the rise in real balances, however, real wealth fell. The effect of an increase in real balances on net worth (emphasized in the Pigou effect) was overwhelmed by the debt problem and the fall in the real value of real assets.

Bernanke draws an important policy conclusion from the destructive effects of the debt crisis. Since he views the debt crisis as an exogenous event, he argues for selective bailouts of bankrupt firms. We find this proposal ill-advised and unnecessary. It is ill-advised because it disregards the serious moral hazard associated with such a policy and the incentives it creates in the political process. It is unnecessary, we believe, because the debt crisis, like the banking crisis, is avoidable if the monetary authority prevents severe price deflation. By preventing deflation, the monetary authority prevents

the destructive effect of the money-credit decline and the wave of bankruptcies. We conclude that banking crises and debt crises can be prevented with the aid of a suitable choice of monetary arrangements.

### III. Loan Rationing

All shocks operating on the economy induce adjustments in portfolios and impose changes in relative asset prices. These changes in relative prices are part of a general substitution process affecting all assets and liabilities. This process transmits the asset markets responses to the output market, reinforcing or moderating any direct effects of the shocks to the output market. In particular, monetary impulses are transmitted to the output market via the general substitution process and resulting relative price changes.

Stiglitz, in a number of papers, seems to contest this analysis. He assigns a central place in the transmission of monetary impulses to loan rationing. This assignment is supported with the observation that variations of the real rate of interest remain comparatively small over the course of a business cycle. Changes in real rates appear to be insufficient and inadequate as a conduit of monetary impulses. Loan-rationing offers, on the other hand, a powerful, if somewhat asymmetric, conduit, since contractive monetary impulses are more reliably transmitted than expansive impulses.

The objections advanced by Stiglitz and others may be relevant in the context of the IS/LM framework, particularly when interest rates are interpreted as borrowing costs. The situation changes, however, when the transmission mechanism includes a spectrum of assets and liabilities. The magnitude of changes in interest rates observed over cycles is sufficient to produce substantial changes in asset prices. Indeed, a major issue in contemporary finance is whether asset prices fluctuate more than can be explained with current models of asset prices (Robert Shiller, 1981). The (impressionistically) moderate movement of interest rates cannot establish that the general substitution process, involving relative prices of real and financial

assets, cannot explain the observed adjustments.

The problem vanished in a multi-asset model. The extended asset market analysis summarized in our papers (1987; 1988) implies that the transmission of monetary impulses via non-Keynesian channels may strengthen under conditions which weaken the "Keynesian channel." Stiglitz's objection reflects the basic inadequacy of an IS/LM analysis which neglects a credit market and its interaction with a money market.

Our critique is addressed to the analysis presented by Stiglitz and his coauthors. It does not affect the relevance of the phenomenon addressed under the label of "loan rationing." The phenomenon would not arise in a Walrasian world with full (or nearly full) information. Loan rationing arises in a world of uncertainty, a world with transaction costs and costs of information.

Banks post a loan rate applicable to lowest-rate customers with small transaction costs. This prime rate is supplemented with an internal schedule specifying a range of loan rates for higher-risk classes and customers with higher transaction costs.

Consider the situation confronting a bank which sets a prime rate reflecting its assessment of market conditions expressed by a range of market interest rates. Every loan application involves risk, potential information, and transaction costs. Investment in information may lower the risk. Risk premiums, information and transaction costs, however, reduce the net loan rate received by the bank below the scheduled loan rate paid by the borrower. The net loan rate guides the bank's decision and the scheduled rate the borrower's decision. The wedge between scheduled and net loan rate is not constant. Applicants with highly uncertain repayment, large information costs required to lower the risk, and potentially large transaction costs are rejected. The expected net loan rate is too uncertain and too low compared to relevant opportunity costs. Moreover, under such circumstances, raising the scheduled rate may raise the implicit risk premium even further. Under pronounced uncertainty, raising loan rates may not be a solution to the bank's problem. A deliberate selection of loans can

solve the bank's allocation problem. Loan applications with expected net real return at least equal to alternative return opportunities are selected for servicing by the bank.

Our analysis implies that the widespread custom of interpreting "nonprice rationing" as a sign of market failure is misconceived. Reliance on allocational mechanisms other than explicit prices characterizes many markets in which uncertainty about major aspects of the relevant product or service has a large role. Loan rationing is one such mechanism. It is not the central arch of the monetary transmission mechanism, as Stiglitz suggests. Once we move beyond the IS/LM analysis by incorporating a credit market and introducing a general substitution process, loan rationing supplements interest rate rationing, and other responses to relative price changes, as part of the monetary transmission process.

## REFERENCES

- Bernanke, Ben, "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review*, June 1983, 73, 257-76.
- Brunner, Karl, "Survey of Selected Issues in Monetary Theory," *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 107th year, No. 1, 1971.
- \_\_\_\_\_, and Meltzer, Allan H., *Money and the Economy: Issues in Monetary Analysis* presented at the Raffaele Mattioli Lectures in Milan, November 1987 (Cambridge University Press, forthcoming, 1989).
- \_\_\_\_\_, and \_\_\_\_\_, "Money Supply," unpublished manuscript, 1988.
- Cagan, Phillip, *Determinants and Effects of Changes in the Stock of Money, 1875-1960*, NBER, New York: Columbia University Press, 1965.
- Friedman, Benjamin, "Monetary Policy with Credit Aggregate Target," *Carnegie-Rochester Conference Series on Public Policy: Money, Monetary Policies and Financial Institutions*, Spring 1983, 18, 117-47.
- Schwartz, Anna J., "Financial Stability and the Federal Safety Net," working paper

American Enterprise Institute, November 1987.

Shiller, Robert, "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?" *American Economic Review*, June 1981, 71, 421-36.

Stiglitz, J. E., "The Causes and Consequences

of the Dependence of Quality on Price," *Journal of Economic Literature*, March 1987, 25, 1-48.

\_\_\_\_\_ and Weiss, Andrew, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, 71, 393-410.

# COMPARATIVE STRATEGIES FOR ECONOMIC REFORM<sup>†</sup>

## Economic Reforms Within and Beyond the State Sector

By TAMÁS BAUER\*

The literature discussing the problems of economic reform in planned economies usually presents the problem as that of transforming the institutions, control methods, and behavioral rules applicable to the state and cooperative (socialist) sectors of the economy. The fact that a certain part of private handicraft, trade, and agriculture survived the postwar nationalization and collectivization drives and was even resuscitated under the New Course (1953–54) or following 1956, has been well-known but neglected in analyses of reform history.

### I. Problems and Concepts

János Kornai (1983; 1986) extended the descriptive concept of economic reforms to changes beyond the state and traditional or large-scale cooperative sectors, reflecting the shift in actual reform policies. While during the 1960's the emphasis was on reforming the state and large-scale cooperative sectors in Hungary, during the early 1980's the growth of the private and semiprivate sectors represented the most important changes in the economic system. The expansion of private agriculture, handicraft, and trade has been the dominant element of change also in China's economic system during the last decade. This paper will discuss the distinction between these two types of reforms and their sequence in planned economies.

Following Kornai, I distinguish between two main sectors in centrally planned econo-

mies. The first is the state sector and the large-scale cooperatives in industry, agriculture, and domestic trade, the dominant part of each East European economy's industry, foreign trade, domestic trade, and construction; it includes also a minor but important part of agriculture. The status of cooperatives that dominate in agriculture and play a minor but important role in retail trade (in some countries also in industry and in construction) is legally different, but their actual operation is very close to those of state enterprises: they are subject to hierarchical planning to the same degree as state enterprises, their managers are subject to the *nomenklatura*, and their self-management system is formalistic. For the sake of simplicity, let this be called the first sector.

The rest of the economy constitutes the second sector, more or less identical with the second economy, in a broad sense of the term. Private agriculture, retail trade, services, and handicrafts are the traditional "nucleus" of this sector. During the last more than thirty years, numerous forms of mixed or semiprivate economy have emerged in all countries: the household plots of collective farmers; the household plots of non-agricultural employees; auctioned-lease operations in retail trade and catering; contract work associations that may be independent or operate within an enterprise, etc.; non-licensed black- or gray-market economic activities, particularly in construction and in repair services, etc. Let the sum of these be called the second sector.

### II. Hungarian Development

In Hungary, the first suggestions to reform the economic system were put forward shortly after Stalin's death, during the New Course years in the mid-1950's. Following

<sup>†</sup>*Discussants:* Ed A. Hewett, The Brookings Institution; Nicholas Lardy, University of Washington; Paul Marer, Indiana University.

\*Institute of Economics, Hungarian Academy of Sciences, H-1502 Budapest, P. O. B. 262, Hungary.

the suppression of the 1956 uprising, the Kádár government convened the best economists of the country to draft a blueprint of policy and systemic changes. A proposal emerged to rebuild the system of economic planning and control in a way very close to the New Economic Mechanism as it was introduced in 1968 (see Iván Berend, 1983; László Szamuely, 1982; 1984).

Following the rapid consolidation of the regime in 1957, the proposed reform program was put aside. The traditional, Soviet-type planning system was maintained, with some minor changes. Most analysts describe this process as abandoning the course of reforms. Berend, while admitting that the 1957 reform proposals were, in fact, rejected, notes that nevertheless the "first reform steps" were taken. In a certain sense he is right. The government made several important concessions concerning economic activities beyond the state sector which, at the time, did not constitute parts of a comprehensive reform blueprint, but were conceived as *ad hoc* concessions. Ultimately, however, they turned out to be important building blocks in transforming substantially Hungary's economy and society.

Private handicraft and retail trade, strongly constrained during the early 1950's, were encouraged during the late 1950's and recovered to a certain degree.

Following a brief interval between late 1956 and early 1959, during which private agriculture flourished, the sector was collectivized between 1959 and 1961, essentially along Soviet lines. But during the mid-1960's, the pattern of collectivized agriculture was substantially reshaped. Household plots acquired much higher importance, and collective farmers secured more freedom of action on their household plots than under the traditional *kolkhoz* model.

The withdrawal of the state from housing construction and that of the peasants from investing in farming resulted in an upswing in private housing construction, first in rural areas and later in urban areas. Most of this construction was done by mutual self-help and by unlicensed building entrepreneurs. The growing share of private housing construction not only enabled people to build

for themselves (or to let private contractors build for them), but also motivated them to earn additional money in the second economy.

At the same time, as the growing demand for services could not be met by the first sector, unlicensed private services were growing.

These changes were not part of the implementation of a single reform blueprint concerning Hungary's second sector. They resulted from concessions the authorities made to remove the most obvious cases of rigidity and irrationality from the economy, to relieve the state budget from the burden of subsidizing housing construction, and to force the population to spend their incomes for items which otherwise would have had to be financed from the state budget. These changes made the operation of the economy somewhat more flexible and less costly (at least for the government; in a certain sense, it became more costly for individuals), but did not eliminate the basic mechanisms responsible for wasting resources that are so characteristic of centrally planned economies which continued to operate in the first sector. The economic tensions of the early and mid-1960's made this clear for the party leadership, prompting it to reform the first sector of the economy, a decision made during 1965-66.

A comprehensive "reform of the economic mechanism" was introduced in 1968. It resulted in the emergence of a peculiar mixed system described extensively in the literature, among others by Kornai (1983, 1986) and by myself (1983, 1986). Mandatory planning and centralized resource allocation were abandoned; prices, wages, and employment were to some extent liberalized; and selected firms obtained direct access to the external markets. This time the emphasis was on changes within the first sector: state enterprises obtained a high degree of autonomy, were relieved from the constraints of hierarchical planning, and, in most cases, were also freed from compulsory membership in trusts or associations. Even greater were the changes in the cooperative sector: many collective farms developed into big agricultural enterprises with a strong busi-

ness-like orientation. The 1968 reform did not imply a substantial shift in the position of the private sector (the trend of slow reduction of private handicraft and trade continued), but nevertheless it gave a new impulse to the development of the second sector, owing to the liberalization of labor legislation and the new status of cooperatives, which made possible the creation of subsidiaries by agricultural and retail trade cooperatives operating in industry and construction. This was the first important form of semiprivate economic activity characterized by strong profit motivation and *de facto* independence from the bureaucracy. (For an excellent analysis, see Kálmán Rupp, 1983.)

The operation of the Hungarian economy became substantially more flexible and rational in comparison with the prereform state of affairs, but remained far from becoming a market economy. Despite the reform, much remained unchanged. The concentration of the bulk of social capital in a small number of large firms and cooperatives, the centralization of the function of initiating innovations, the central control over a large part of investment, the wide-scale manipulation of prices, taxes, and subsidies, and the frequent interventions in commodity transactions proved to be sufficient to prevent a marked improvement in macroeconomic performance.

After the period of "freezing the reform" during the early and mid-1970's, acute economic tensions developed, forcing the authorities to resume the policy of reform in 1979. The reform proposals encompassed both the first and the second sectors. Numerous decrees and new laws were issued by the government which changed substantially the status of state enterprises: direct supervision over the majority of state firms by the sectoral ministries was abandoned, foreign trade monopoly was weakened, elements of a capital market emerged, and the monobank was replaced by a two-tier banking system consisting of a central bank and several competing commercial banks. Still, the evidence supports Kornai's conclusion that little has changed in the position of the state firms (and of the large-scale cooperatives) regarding their dual dependence on the

authorities and on their customers. With respect to the first sector, the new measures were in fact ineffective, because some basic constraints on competition (for example, the degree of monopolization of the economy, import and price controls) were maintained, and in some respects even reinforced. The suggestions of economists to go further in the direction of creating a real market mechanism by introducing a "package" of simultaneous reforms were rejected.

At the same time, the measures aiming at the development of the private and particularly of the semiprivate sector proved to be much more successful. In the early 1980's, the authorities adopted a comprehensive program to develop new forms of business. In a few years, more than 10,000 small new enterprises emerged (in the forms of "petty cooperatives," "independent contract work associations," etc.). In a country where there were fewer than a thousand state and cooperative enterprises, this was quite an achievement. Private retail trade and auctioned-lease operations also expanded rapidly. The expansion of the second sector contributed considerably to maintaining a relatively good supply situation on the domestic market. During the 1980's, the entire increment in national product was contributed by this sector.

In a certain sense, the period of the 1980's was a repetition of the late 1950's and early 1960's: a substantial enlargement of the second sector without (or instead of) reform in the first sector.

### III. The Polish Case

Poland is the only other East European country where similar trends in reform policy can be observed. The blueprint of a comprehensive reform was drafted by a similar commission of experts, and several important suggestions of the commission were implemented during the late 1950's. However, the systemic innovations were inconsistent, had very limited impact on the operation of the system, and were shortly withdrawn. At the same time, substantial concessions were made regarding the private sector: the collectivization in agriculture was



abandoned and private handicraft and retail trade underwent a revival.

Without much fanfare, during the early and mid-1970's, private activities became subject to more constraints. Interestingly, this coincided with the WOG experiment that represented an effort to reform the state sector (which proved to be unsuccessful). During the late 1970's and early 1980's, new concessions were made to the semiprivate and private sectors (for example, Polonia firms, interenterprise contract work associations, more guarantees to private agriculture). The equal treatment of different sectors was proclaimed by the Polish as well as the Hungarian authorities. (The history of the nonagricultural private sector in Poland is discussed in Anders Åslund, 1985).

#### IV. Reforms in their Sociopolitical Context

Let me draw attention to the relationship between the two types of reforms described above and their sociopolitical environments, focusing especially on the kinds of resistance they face.

In both cases, the reforms are in conflict with official ideology. The reforms in the first sector call into question mandatory planning as a necessary feature of a socialist economy and the interpretation of socialism as an economy without markets. This, however, did not present a serious problem because the mid-1960's were characterized by an ideological "renewal" in Eastern Europe: the traditional Stalinist picture of socialism was replaced by a new vision of a decentralized, competitive, self-managed, humanized socialism, not inconsistent with the kinds of reforms being introduced in the first sector.

Concessions to the private and semi-private sector call into question the socialism as negation of private ownership and private entrepreneurship. Yet periods of such reforms (the late 1950's and the 1980's) coincided with a forced pragmatism in economic policy, with a turn away from ideological orthodoxy, so that ideology did not present an insurmountable obstacle in the reforms.

More troublesome is the fact that the reforms imply changes in income distribution. The reforms oriented toward the first sector

disturb income distribution among industries and occupational groups and promote intersectoral labor migration. Concessions to the second sector enlarge new sources of personal incomes, new types of careers, which are in sharp contrast with the income distribution principles of a Soviet-type economy. The income hierarchy characteristic of such systems and the ability of the authorities to control it are hurt in both cases.

Even more problematic is the disturbance reforms create for vested interests. Concessions to private initiative and private business at the periphery of the economy conflict with the interests of local (party and government) authorities which lose part of their control over local markets. Also, state enterprises in trade and services now face new competitors. Reforms in the first sector, in turn, hurt the interests of such strong power groups as the "captains of industry" of big firms and the officials in the industrial ministries, in the planning bureaucracy, and in the regional party and government bodies. The resistance of such groups has proven to be much stronger than that of the first type of interest groups.

It is this latter aspect that seems to explain why reform policy has often turned toward concessions to private initiative instead of pursuing vigorously the proclaimed reforms in the first sector. In China, too, reforms in the second sector have been much more effective than in the first sector.

The course of *perestroika* initiated in the USSR and followed also, at least in rhetoric, by Bulgaria and Czechoslovakia implies changes of both kinds. The scope of private initiative is being extended and new legislation relative to the status of state enterprises is being passed. Considering the sociopolitical dynamics of economic reforms in Hungary and Poland, it seems quite likely that a course of events similar to that in the two countries during the late 1950's and early 1960's may develop in countries like the Soviet Union and Bulgaria also.

In Hungary and Poland, further concessions to private initiative can contribute less and less to a substantially improved national economic performance under an unchanged system in the first economy. The losses

originating in the first sector have grown rapidly enough to offset much if not all of the contributions of the second sector to national welfare. A shift toward genuine reforms of the first economy, creating conditions for genuine market competition and thus eliminating loss-making firms and encouraging the profitable ones, is urgent. It remains an open question whether such essential further reforms can be implemented.

#### REFERENCES

- Åslund, Anders, *Private Enterprise in Eastern Europe*, London: Macmillan, 1985.
- Bauer, Tamás, "The Hungarian Alternative to Soviet-type Planning," *Journal of Comparative Economics*, September 1983, 7, 304-16.
- \_\_\_\_\_, "Reforming or Perfecting the Economic Mechanism in Eastern Europe," Working Paper No. 86/247, European University Institute, Florence, 1986.
- Berend, Iván T., *Gazdasági Utikeresés 1956-1965* (Searching for an economic path 1956-1965), Budapest: Magvető, 1983.
- Kornai, János, "Comments on the Present State and Prospects of the Hungarian Economic Reform," *Journal of Comparative Economics*, September 1983, 7, 225-52.
- \_\_\_\_\_, "The Hungarian Reform Process: Visions, Hopes and Reality," *Journal of Economic Literature*, December 1986, 24, 1687-737.
- Rupp, Kálmán, *Entrepreneurs in Red*, Albany: State University of New York Press, 1983.
- Szamuely, László, "The First Wave of the Mechanism Debate in Hungary (1954-1957)," *Acta Oeconomica*, Nos. 1-2, 1982, 29, 1-24.
- \_\_\_\_\_, "The Second Wave of the Economic Mechanism Debate and the 1968 Reform in Hungary," *Acta Oeconomica*, Nos. 1-2, 1984, 33, 43-67.

# On the Strategy for Implementing Economic Reform in the USSR

By VALERY L. MAKAROV\*

The economic reform under way in the USSR is part of a general process of *perestroika* (restructuring), encompassing all spheres of social life. Consequently, the reform measures discussed below should be understood and analyzed in the wider context of that general process. Nevertheless, I will only be able to focus here on those parts of *perestroika* directly concerning the economy.

Our reform strategy can be divided into three phases. Phase I (1985–87) has focused on the understanding of the past and the development of a strategy. Phase II (1988–90) will be a transitional phase during which the new system will be introduced, while many features of the old system will linger. Phase III (1991–95) will be the first full five-year plan period in which the entire system will be in place and functioning.

## I

Phase I, which began with Mikhail Sergeevich Gorbachev's elevation to the post of General Secretary, was characterized by an effort to understand the previous period of economic development, and to develop the guiding principles for the new reform. Throughout this period the trend was in the direction of increasingly radical proposals and measures as the understanding developed that the reform would have to be radical, changing the very roots of the existing economic system. Evidence of the increasingly revolutionary nature of the proposed economic changes can be found, for example, in a comparison of Gorbachev's June 1985 speech to a Central Committee meeting on science and technology, and his speech to the June 1987 Central Committee Plenum.

At the same time, some economic authorities, including some enterprise managers and government officials in the planning organs, expressed doubts concerning the need for such radical changes. And these doubts showed some tendency to spread. The increasingly radical nature of reform formulations was, in part, an effort to extinguish, or at least neutralize, the spread of these conservative views.

The experience in 1986 with the first reform decrees also served to emphasize to the leadership the necessity of more radical measures. In 1986 joint party-government decrees were adopted concerning changes in the economic framework of the agro-industrial complex, light industry, and in other areas. It very quickly became apparent that these measures were too modest to bring about radical changes; the fundamentals of the system remained unchanged after the decrees were implemented.

Our experience during this period led to a growing conviction that a reform could only succeed if it was designed according to an overall strategic plan encompassing a comprehensive set of measures involving carefully coordinated reforms in planning, price formation, financial-credit institutions, material-technical supply, and so on. The result of this consensus is evident in the "Basic Provisions for the Radical Restructuring of the Management of the Economy" ("Osnovnye polozheniia...", 1987) approved at the June 1987 Central Committee Plenum, and the "Law on State Enterprise" ("Zakon...", 1987) adopted by the Supreme Soviet at the end of June. In addition, the basic decrees guiding the implementation of the reforms were issued as a package in July 1987 (*O korennoi perestroike...*, 1987).

## II

We are now entering Phase II of the reforms, which will run approximately through

\*Director, Central Economics-Mathematics Institute, USSR Academy of Sciences, Moscow.

1990. This phase is decisive to the process of *perestroika*, for it is here that the actual nature of the reforms will be determined.

One of the important "givens" of this phase is the Twelfth Five-Year Plan (FY-PXII), which was approved in 1986. No one can cancel this plan; it is the law, serving as a guide to ministries and enterprises in their daily supervisory and productive activities. That creates a contradiction during this transitional period. There is, on the one hand, an approved plan. Yet, on the other hand, there is a new law on the socialist enterprise, according to which the enterprise will construct its own plan.

The contradiction has been resolved by converting the targets of FYPXII into state orders (*goszakazy*). In the reformed system, state orders are intended as a contractual agreement between Gosplan, the ministries, or the republican authorities, on the one hand, and the enterprise on the other, with mutual obligations—in other words, an economic contract, freely entered into by both sides. And the intention is that state orders will not encompass anywhere near all of national output; rather they will be limited to ensuring supplies for the most important projects in the social sphere, state investment projects, and defense. But the use of state orders during this second phase as a vehicle for conveying FYPXII plan targets to enterprises, departs from the principles of the reform, continuing the old system of plan commands to enterprises.

Further contradictions are apparent in the way that central authorities are applying "norms" during this transition phase. In the new system, enterprises are to be given a set of *normativy* (normatives) specifying 1) the share of profits going to the state, the ministry, and the enterprise; 2) the authorized size of the wage fund (linked to performance of the enterprise); and 3) charges for working capital, the use of natural resources, and so on. These normatives are to be fixed for five-year periods (with the exception of FYPXII in which the norms were set for 1988–90), so that enterprises will know there are hard and fast rules of the game, which will apply to all enterprises, whether they perform well or poorly.

However, because the FYPXII targets remain in force, the normatives have been tailored to each enterprise for the 1988–90 period, a direct contradiction with the spirit of the reforms. This problem of normatives will emerge with full force in this phase of the reforms; and much remains unresolved, requiring further work and discussion.

One of the more interesting features of Phase II is the adoption in the civilian industries of techniques successfully applied heretofore only in the defense sector. The most obvious example is the close copy of military quality control procedures (*gospriemka*) applied in selected civilian enterprises during 1987 which will be introduced more broadly in 1988. The introduction of a *biuro* to oversee the civilian industrial ministries, similar to the Military-Industrial Commission, is another example. Finally, a significant portion of the economic leaders in important positions in civilian industry have backgrounds in military industries (Nikolai I. Ryzhkov, the Prime Minister; Ivan S. Silaev, the head of the Machine-building *biuro*; and Lev A. Voronin, Chairman of Gossnab).

These measures may have some positive short-term effects since the military sector of our economy has historically been far more efficient than the civilian sector. But these are nevertheless administrative measures, and however effective they may be in the near term, in the long run they will contradict the reform.

During this second phase we will also begin the implementation of reforms in price formation and in the credit-financial system; the widespread introduction of wholesale trade will commence; and scientific research institutions will be transferred to full cost accounting. Mixed enterprises will be created, including Soviet-Western enterprises; individual and cooperative activity will increase.

The implementation of such a reform in a three to five-year period is a revolutionary, not an evolutionary, process. Inevitably there will be winners and losers. The losses will focus particularly on those strata of society whose services will be in far less demand in the new system. Government bureaucrats,

whose numbers will be drastically reduced, will lose their economic and social positions. A significant number of workers in industry and agriculture will also be affected as they discover that they must work harder to retain their old incomes and living standards.

The essence of these revolutionary changes is a transition from an excessively stable and rigid economic system to one which is much more flexible. We must move towards a system in which it is far more common, and easier than it has been in the past, to change one's place of work and residence, to open and close enterprises, introduce new innovations, to replace old with new products. Changes of such magnitude are enormously difficult, and surely cannot be fully achieved in a mere three to five years. But we must begin the process now; there are no alternatives. If we wish to compete in the world economy, where technological change is rapid and constant, we must move as quickly as possible to a much more flexible economic system.

All of this will lead to a mixture of the old system, with the new; of old rights, instructions and decrees, with new ones. As a result we should expect to see in selected regions or sectors a decline in growth rates, and other economic indicators. It is conceivable that selected strata of the population will experience a temporary drop in their living standards in some regions. Because of this the second phase of the reforms will be politically difficult. We will need to carefully explain to the population precisely what we are doing, and to prepare them for possible negative consequences of reform measures.

### III

Phase III, during the first half of the next decade, will see the transition to a relatively stationary economic regime. The Thirteenth Five-Year Plan will be formulated according to the rules of the new system. Enterprises will receive tasks from above only in the form of state orders, and the share of state orders in total output will fall year by year. The new price system will be in place; capital goods and intermediate products will generally move through wholesale trade channels.

The bulk of capital expenditures will be financed directly out of enterprise earnings, or through bank credits. By the beginning of this phase of the reform, enterprise leaders should have overcome the psychological barrier connected with the veneration of higher organs, having learned to make their own decisions concerning commercial activities.

These measures will combine to bring the full force of market-type mechanisms to work on the economy. The dispersion of enterprise and work income will begin to grow, a function of their economic performance. That, in turn, will lead to a wider dispersion of income in the population as a whole.

One of the critical preconditions to the success of this third phase will be a change in the psychology of working people regarding high incomes. The fact is that, for a long time, workers and many government bureaucrats have come to the conviction that the desire for high incomes is a bourgeois relic. And, in essence, this had become the official point of view. People who are receiving significantly more than others in a particular work situation generally do not enjoy the support of public opinion. A change in that viewpoint in favor of higher prestige for those who have truly earned higher incomes will be critically important to improving the material stimuli for harder work.

As a result of the full implementation of the economic reform in this third phase, the economic indicators should gradually improve as improvements in factor efficiency lead to higher growth rates. We will also see during this period a dramatic expansion in the activities of small units, including those operated by cooperatives and individuals, with a particular focus on services, and other areas connected with personal consumption.

In sum, by the end of Phase III, when these measures are fully implemented and have been working for some time, we should see a significant improvement in consumer welfare.

### REFERENCES

- Gorbachev, M. S., "Korennoi vopros ekonomicheskoi politiki partii. Doklad tova-

rishcha M. S. Gorbacheva" (The Fundamental Issue of the Economic Policy of the Party. Report of Comrade M. S. Gorbachev), *Pravda*, June 12, 1985, 1-2.

\_\_\_\_\_, "O zadachakh partii po korennoi perestroike upravleniia ekonomikoi. Doklad General'nogo sekretaria TsK KPSS M. S. Gorbacheva na Plenum TsK KPSS 25 Iiunia 1987 goda" (On the tasks of the party for the radical restructuring of the management of the economy. Report of the General Secretary of the CC of the CPSU M. S. Gorbachev to the Plenum of the CC of the CPSU, June 25, 1987), *Pravda*, June 26, 1987, 1-5.

*O korennoi perestroike upravleniia ekonomikoi* (On the Radical Restructuring of the Management of the Economy), Moscow: Politicheskaya literatura, 1987.

"Osnovnye polozheniia korennoi perestroiki upravleniia ekonomikoi" (Basic Provisions for the Radical Restructuring of the Management of the Economy), *Pravda*, June 27, 1987, 2-3.

"Zakon Soiuza Sovetskikh Sotsialisticheskikh Respublik. O gosudarstvennom predpriatii (ob "edinenii)" (A Law of the Union of Soviet Socialist Republics. On the State Enterprise [Association]), *Pravda*, July 1, 1987, 1-4.

# Choosing a Strategy for China's Economic Reform

By JINGLIAN WU AND BRUCE L. REYNOLDS\*

Ever since 1956, although the Chinese reform debate in each period has had its own special characteristics, the argument over choice of strategy has always revolved around two approaches. The first approach usually makes the following points. First, overconcentration of decision-making power is the main defect of the traditional socialist economic system—an excessive restriction of the initiative of local governments and producers. Second, any measure that serves to break up this overconcentration and to stimulate local governments, enterprises, and individuals is an appropriate part of reform. Third, in achieving this stimulus, any and all reforms, whether designed to delegate power or to strengthen material incentives, should be supported. According to the second approach, by contrast, the main defect of the old system is that it allocates resources through administrative commands. Such a system cannot use resources efficiently. And the only system which can substitute for administrative commands in allocating resources is the market mechanism.

Early in 1956, the leaders of the Party decided that China must undertake reform of the economic management system. That decision grew out of criticisms of the way in which the traditional system had operated during the first Five-Year Plan (1953 to 1956). The understanding prevailing at that time can be seen most clearly in the speech, "On the Ten Great Relations," by Mao Zedong, at a meeting of the Political Bureau of the Chinese Communist Party in 1956. At that time, some economists advocated market-oriented reforms (see Zhun Gu, 1957). But this argument had little impact on

Chinese economists, because most of them were still constrained by traditional socialist economics. Based on this majority view, Mao indicated that the direction of economic reform should be to enlarge somewhat the powers of local versus the central government, and to give every production unit more independence in relation to the state. At the same time, the leadership should be concerned for the living standard of the masses, raising their incomes as output rose.

Mao's ideas about decentralizing power and strengthening material incentives were later revised. In the wake of the Anti-Rightist Campaign in 1957, material incentives and enterprise self-management were both regarded as revisionist. The subsequent reform did not stress enlarging the decision-making powers of enterprises or raising the material benefits of workers and peasants, but instead implemented what has been called "decentralization I" (H. F. Shurman, 1968, pp. 175–78) or "administrative decentralization" (see Morris Bornstein, 1977).

Following this line of thought, the 1958 economic reform mainly enlarged the decision-making powers of local governments over raw materials supply and investment. Specific changes included the following. 1) The great majority of central ministry enterprises were transferred to local control. 2) In production planning, the system of unified balance achieved through the State Planning Commission was transformed into one based on regional planning and bottom-to-top balancing. The number of products under the control of the State Planning Commission was greatly reduced, and local authorities were granted important planning and allocative powers. 3) In investment planning, a subcontracting system was adopted: the central authorities distributed funds to the local authorities, who added their own funds and then chose investment projects on their own. 4) Reforms reduced the amount and the categories of materials inventories dis-

\*Senior Advisor, Economic, Technical and Social Development Research Center, State Council, Beijing, PRC, and Associate Professor, Department of Economics, Union College, Schenectady, NY 12308, respectively.

tributed by the State Planning Commission and the ministries, leaving the rest to be distributed by local government. The local authorities could reallocate centrally rationed materials in the hands of enterprise in their region, including central ministry enterprises, and could also share in the above-plan output produced by local enterprises. 5) The financial system was sharply decentralized: different taxes were assigned to different government levels for collection, with various tax retention rates, each fixed for five years. Meanwhile, the central government transferred the power to reduce, exempt or increase a tax to the local authorities. 6) The former highly centralized credit system was replaced by delegating credit control to lower levels and controlling only the debit-credit differential.

At the same time, the central authorities extended enterprise decision-making power. They greatly reduced the scope of mandatory plan targets. The proportion of earnings that an enterprise could retain, which previously was uniform within a given industry, now would vary from enterprise to enterprise. They expanded the enterprise's powers to manage their own personnel and their institutional structure. Lastly, a portion of enterprise capital would now be allocated by the enterprise; moreover, the enterprise could reduce, increase, or even dispose of fixed capital.

These two sets of decentralizing measures were vividly described in a recent book as "blindly delegating administrative power" and "expanding enterprise autonomy amid chaotic macro-management" (Taihe Zhou et al., 1984). They provided the institutional basis for the Great Leap Forward, and were the main cause of the economic chaos in 1958.

After the failure of the Great Leap, China's government took a series of recentralizing measures. However, the intended recentralization has never been completely achieved. China's local governments have more power than in other command economies, and weak restraint by the central authorities has become ingrained, creating what some call a bargaining economy (see Gene Tidrick, 1986).

This is not to say that China abandoned administrative decentralization after 1958. Allocative inefficiency, the inherent defect of any command economy, continually brings reform back onto the agenda, and administrative decentralization always appears to be the only reform option. And so in China, after 1958, similar reform policies were adopted again and again. For instance, the large-scale economic reform launched in 1970 propounded the slogan: Delegating powers to lower levels is a revolution; the more you do, the more revolutionary you will be (Zhou et al., pp. 134-46). Rather than learning the appropriate lesson from the 1958 experience, China settled into the familiar "reform cycle": "We relax control, and get chaos; we recentralize, and get inertia."

In December 1978, after twenty years of silence, a reformist spirit once again burst forth in China, at the Third Plenum of the CCP. The Plenum made clear that Mao's 1956 report was fundamentally correct, by stressing the need to decentralize, stimulate initiative, and act according to economic laws and the law of value. The main content of this reform was, first, to delegate decision-making power to local authorities and production units, and second, to allow local authorities and enterprises to retain more revenue, so as to stimulate their initiative. This approach as summarized in the slogan "fangquan rangli," or "Delegate Power and Relinquish Revenues" (hereafter DPRR).

The basic spirit of reform after 1979, in its advocacy of DPRR, was close to that in 1956. But there were major differences in implementation: 1) The stress in the 1958 reform was on administrative decentralization; reform after 1979 laid more emphasis on the expansion of enterprise decision-making power. 2) The 1958 reform was carried out mainly in the state-owned industrial sector. Reform after 1979 had a much wider scope; in particular, it was implemented in agriculture and foreign economic relations. These differences helped the 1979 reform score achievements that could never have been gained by the reform of 1956-58.

The post-1979 reform consisted of four main changes. First, two major measures were adopted in agriculture. One was a sharp



rise in the state purchase price for farm products. In addition, the state purchase quota was reduced. As a result, peasant income increased. The other measure was to replace the People's Commune, which featured unified management of land and unified distribution of income, with the household-based contract responsibility system, by which land was leased to peasants for up to fifteen years. These two measures quickly stimulated peasants to increase production and to provide good management. Second, the urban and rural collective and private economy was encouraged to grow rapidly. From 1979 to 1984, there was a 14.52 million increase in urban collective employment. Urban and rural private sector jobs grew from virtually zero to 3.4 million. Rural non-agricultural employment (private and collective) now provides more than 80 million jobs, in a total rural labor force of 370 million. Third, in the area of center-local relations, in addition to continuing to share enterprise managerial power with localities, the highly centralized budgetary system was changed to one featuring stratified management with contractual revenue sharing ("eating in separate canteens"). The new regulations fixed revenue-sharing ratios for five years, and local governments may use their surpluses without approval from higher authorities. And fourth, the post-1979 reforms included a major, dramatic opening to the international economy and the outside world.

In these areas, reform scored major successes. But within the state-owned industrial sector, with which this paper is primarily concerned, the picture was somewhat different. Here, the essence of DPRR-type reform was to enlarge enterprise self-management powers and to create an incentive mechanism based on profit sharing between state and enterprise. These powers were soon extended to 6,600 state enterprises, accounting for 60 percent of state-sector output value and 70 percent of total profit. These enterprises did in fact display increased initiative. But, because they were not constrained by a competitive market context (since parallel reforms, especially price reform, had not been carried out), their actions did not necessarily

conform to the good of the national economy. In particular, pressure to increase investment ultimately led to generalized excess demand.

These disadvantages of DPRR reform were identified very early. In 1980, Xue Muqiao argued that price reform and goods circulation deserved the greatest attention, and suggested abolishing administered prices and introducing competitive commodities and financial markets. But in 1981, as problems precipitated an economic readjustment, some officials blamed the difficulties on too much stress on market relations, rather than on the absence of a market-based macro-control system. And so new rules constricted enterprise operations, and those who supported market regulation were criticized (Wu and Renwei Zhao, 1987).

But reform regained momentum in 1984, and in May, the State Council promulgated the Provisional Regulations Concerning Further Extending the Decision-Making Power of State-Owned Enterprises. State-owned enterprises were assured the following powers: 1) to set production according to market demand, after fulfilling state plans; 2) to sell, on their own, a certain part of their production; 3) to set prices on the products they sell on their own; 4) to utilize the profits retention fund; 5) to lease or sell idle assets; 6) to rearrange their staff and to appoint middle-level administrative staff; 7) to choose the form of wages and the disposition of bonus funds from retained profits; and 8) to join sectoral and trade associations. Later, in August 1984, the State Council approved Provisional Regulations on Improving the Planning Structure, which reduced the scope of mandatory plans relative to guidance plans and market regulation.

The above reforms helped state enterprises to move away from mere passive fulfillment. There has been a clear reorientation toward the pursuit of profit. Enterprise management is more active. The reforms also invigorated the economy, as can be seen from the rising industrial growth rate (10, 14, and 18 percent in 1983, 1984, and 1985). But improved efficiency did not contribute much to output growth. Our calculations show that, between 1981 and 1984, the annual total factor pro-

ductivity (TFP) growth rate was 0.6 percent, higher than the 0.1 percent for 1956 to 1979, but much lower than the 3.8 percent for 1953 to 1957. In the period 1981–85, the share of increased TFP in the sources of industrial growth was 8.2 percent, lower than that of the major industrial countries. Meanwhile, both the inflation rate and the budget deficit were rising, especially in 1984, when the annual growth rate of currency ( $M_0$ ) reached 50 percent.

Chinese economists have drawn very different conclusions from this ambivalent situation. Some reformers believe that fast GNP growth shows the wisdom of “delegating powers.” This strategy, they argue, has caused the Chinese economy to “take off.” They also believe, pointing to international experience, that a money supply growth rate exceeding that of output value is normal and contributes to economic growth with little risk.

Others, called coordinated reformers (the Chinese co-author is in this group) have a quite different understanding. They argue that the economic results in late 1984 underscore the shortcomings of a DPRR approach. The rapid inflation signalled excess aggregate demand—a poorly functioning economic system with weak macro control. This poor functioning stems, first of all, from a doubly distorted price structure. Thirty years of administrative intervention, in and of itself, has produced irrational relative prices, and the reform-linked dual price system has gone further, creating multiple and widely varying prices for one and the same item. Secondly, decentralization along administrative lines has fragmented national markets and fostered regional protectionism. And in addition to the inflation, other problems also cause concern. Dysfunctional enterprise behavior persists—managers exhibit short time horizons, and are at least as attentive to their administrative superiors as they are to market forces; and rampant profiteering and related problems of income distribution have also arisen.

A planned market economy, says this group, must include three central elements. Enterprise decision-making power, coupled with material incentives, is only one of these.

In addition, markets must be competitive, to generate correct price signals and to facilitate resource reallocation. And third, a system of indirect macroeconomic management must be built up, operating through competitive markets. In creating these three elements, price reform, which to date has been the weakest link in China’s reform, should instead be the first and key step, implemented in coordination with reform of the taxation, fiscal and financial systems (similar to the ideas of John Fei and Reynolds, 1988). And the essential prerequisite for price reform is to largely eliminate excess demand.

By the end of 1984, then, many aspects of the old command economy had been broken down, but the necessary institutional replacements were incomplete. Swollen demand and economic overheating in the last quarter of the year suggested the clear possibility that neither the old nor the new system would work effectively. Furthermore, DPRR required that the state, especially the central government, increase expenditures just as tax and other revenue was shrinking. China was confronted by a massive budget deficit and great inflationary pressures. The situation had to be rectified at once; otherwise, there would be no alternative but to revert to the traditional administrative devices.

During the course of this experience, the ideas of the coordinated reform school won increasing acceptance. They were confirmed by the Party’s National Conference, in October 1985, and were embedded in the CCP Proposals for the Seventh Five-Year Plan. The main direction of economic policy for 1986, as set forth in the Proposals, would be to restrain demand, increase supply, and pave the way for decisive reform measures in 1987 (*People’s Daily*, January 14, 1986). In the meantime, a group of policy advisers was brought together to work out a coordinated reform scenario, to be carried out step by step during the Seventh Five-Year Plan period.

Their proposal, fashioned between March and August, involved four elements. 1) Most importantly, price reform. Initially, administrative adjustments would rationalize relative prices for production materials, energy and transportation; in the second or third

year, production materials would be released to market forces. 2) Tax reform. To create a level playing field for interenterprise competition, the turnover tax would be transformed into a value-added tax, and a set of natural resource taxes (land use and mineral deposit fees) would be introduced. 3) Fiscal reform. The new land use fees would flow to provincial and (through them) to municipal governments. Revenue from enterprise profits (for example, tobacco, alcohol) would be replaced by an excise tax on these goods, and this revenue would flow to the central government. Other taxes (VAT) would be shared among the different government levels. In this way, local and central government revenue would no longer hinge on control of specific industrial enterprises. 4) The reform plan envisaged a refashioned financial and monetary system, within which both commercial and investment banks would be formally independent of government. It was anticipated that with these reforms in place, China would be able, by the end of the Seventh or the beginning of the Eighth Five-Year Plan, to make the market mechanism, under macro regulation, play a leading role in the national economy.

This scenario received the approval of the highest-level authorities. But in the end, it was not put into effect. The reasons were twofold. First, the necessary environment for comprehensive reform—a more or less balanced economy—had not been created. The coordinated reformers generally supported a policy of economic stabilization and strong macroeconomic control (Wu, 1985). But to maintain such a policy is not an easy task. More and more people opposed a tight macroeconomic policy, especially after the industrial growth rate suddenly dropped to 0.9 percent in February 1986. And when the central bank loosened credit in March 1986, the money supply began to increase, starting in the second quarter, at roughly a 25 percent annual rate, greatly surpassing the 1986 national income growth rate of 7.4 percent. All this foreshadowed impending inflation in 1987. Therefore, even the designer of the coordinated reform scenario regretfully concluded that it could not be implemented in 1987.

Second, a challenge arose to the basic principles of coordinated reform. Some reformers held that China's new economic system would emerge from a natural evolution, and that to design a strategic plan beforehand was impossible; what people should do was to reform what could most easily be reformed. There had especially been opposition to price reform. This reform, so the argument ran, involves a fundamental readjustment of vested interests. This cannot benefit every social group; therefore it is too difficult and too risky. Furthermore, two-track pricing, which already exists within the price system, enables the market system to play its function; no further change is needed now. The main problem in China's national economy, they argued, are the vague boundaries among property rights in the state-owned sector, and the resulting diffusion of enterprise management power. Therefore price reform should be postponed, and priority should be given to ownership reform, as this school terms it: creating the microeconomic foundations of the national economy. (This viewpoint is presented in the writings of Li Yining, *Beijing Daily*, 5/19/86; *Lilun Xinxin Bao*, 11/3/86; *World Economic Herald* 11/8/86; and Hua Sheng et al., *World Economic Herald*, 1/19/87.) And furthermore, this group argued, a balance of aggregate demand and supply, which according to the coordinated reformers would have to precede comprehensive price reform, is on the contrary the very goal of reform; it can surely not be achieved until reform nears its conclusion. In a developing country like China, in the predictable future, the national economy will necessarily experience excess aggregate demand. If we try to suppress demand and limit the money supply through artificial macro control, it will damage not only high-speed economic growth, but also the interests of many social groups, and as a consequence, will reduce the people's support for reform.

In the last quarter of 1986, these opinions became dominant, and the coordinated reform scenario was abandoned. The emphasis shifted to "micro reform," taking "chengbao zeren zhi," the contract system in enterprises, as its basic form. As the people who

## APPLICATIONS OF INTERNATIONAL COMPARISONS OF PRICES AND QUANTITIES<sup>†</sup>

### What We Have Learned about Prices and Quantities from International Comparisons: 1987

By ALAN HESTON AND ROBERT SUMMERS\*

We provide a sampling of findings derived from the International Comparison Project (ICP) benchmark studies of the last 15 years (see our 1982 book with Irving Kravis and references therein, and United Nations, 1986). The research results take a variety of forms, and are at various levels of aggregation. We begin with a discussion of *real* gross product and price comparisons, and then go on to more detailed analytical comparisons. One ultimate use of the benchmark (cross-country) comparisons is to combine them with country national accounts (time-series) data to get a System of REAL National Accounts (SRNA). Such a companion to the presently available (SNA) social accounts, would make direct interspatial comparisons possible as well as the present intertemporal ones. In addition, the benchmark studies contain detailed information at a much finer level of aggregation than the national accounts which can be useful to researchers in a variety of ways.

#### I. Gross Domestic Product

While the concept and measure of the gross domestic product (GDP) fail to capture many aspects of life (for example, humanistic values) that are important to socie-

ties, most economists would agree that the level and movement of GDP do tell us a great deal about economic conditions in a country. What exactly do we learn from GDP numbers?

First, consider the GDP figures for a number of countries in 1980, each denominated in its own currency units as they are normally carried in the archives of the United Nations or the World Bank. Of course, no comparative judgments can be made about the countries until these 1980 "facts" are made commensurate by being restated in a common currency unit and then recast in some illuminating way. (The first fact freshmen in an elementary economics course are disabused of is that facts speak for themselves.)

A country's domestic-currency GDP typically has been converted into U.S. dollars, the United States being the usual numeraire country, in either of two ways: using an estimate of the purchasing power parity (PPP) of the country's currency with respect to the dollar, or using the exchange rate the country's currency trades at relative to the dollar on foreign exchange markets. The PPP way is the right way to go, by definition, because PPP is the ratio of the domestic cost of buying a bundle of goods and services in the country at its own prices to the corresponding cost in dollars of the same bundle in the United States. (Left undescribed here is how one deals with the heart of the index-number problem underlying international comparisons: What commodity bundle should be used in calculating each country's PPP?) The use of the exchange rate is sanctioned theoretically only by the absolute version of the Casselian purchasing power

<sup>†</sup>*Discussants:* Jagdish Bhagwati, Columbia University; Christopher Clague, University of Maryland; Dale W. Jorgenson, Harvard University.

\*University of Pennsylvania, Philadelphia, PA 19104. The NSF provided financial support for this research under grant SES-205827. The capable research assistance of Ju Yong Park and Joon Haeng Lee is gratefully acknowledged.

parity doctrine, but that doctrine has long since been discredited.

Still, choices are made on the basis of tradeoffs. Good estimates of PPPs are hard to come by while exchange rates are easy to get. Pragmatically speaking, the exchange rate's usefulness as a proxy for the PPP can be judged by seeing how close exchange-rate-converted dollar values are to corresponding PPP-converted ones. If the PPPs estimated by the ICP are used as the standard for the comparison, the overwhelming evidence is that the exchange-rate-converted numbers differ from the PPP-converted ones *significantly* and in a systematic way. In 1980, out of 60 country comparisons, the ratios of the PPP-derived numbers to the exchange-rate-derived ones vary from .76 (Germany) to 4.4 (Sri Lanka). All but Canada of the 12 richest countries (excluding the United States, that has a ratio equal to unity by definition) had ratios below unity, and all but 6 of the poorest 47 countries had ratios greater than unity. The explanation for *why* the ratio on average is inversely related to country income has been investigated extensively (see Kravis and Robert Lipsey, 1988) and will not be reviewed here.

It would be nice if the exchange rate could serve as a satisfactory proxy for PPP, because then international comparisons would be much easier to carry out. Exchange rates are easily observed but PPPs can only be estimated with great difficulty and at great expense through extensive and detailed price comparisons. (Academics considering direct involvement in international comparison benchmark work should be warned that it is only for the strong-stomached with tenure and deep pockets.)

Before going on to specific uses of GDP numbers, two comments should be made about the combining of benchmark studies results and national accounts data in constructing an SRNA. First, the ICP's various benchmark studies have covered only about half the countries economists are interested in. Short-cut procedures have been devised to estimate real GDP (and consumption, investment, and government) values for non-benchmark countries. Our own previous short-cut procedure (1984) exploited the in-

verse relationship described above, but now we find we do as well by drawing on pricing surveys conducted in major cities worldwide to determine post allowances for international civil servants and businessmen stationed abroad. Second, since benchmark estimates of a country's standing relative to the United States in each of 2 years is likely to imply a different growth rate between the years from the one embedded in the country's national accounts, a reconciling procedure has been adopted to "consistentize" the benchmark and national accounts estimates. (Happily, at the GDP level, the benchmark data reconcile well for rich countries and pretty well for poor ones. Unhappily, they reconcile much less well at the disaggregate consumption, investment, and government levels.)

#### A. *Welfare vs. Productivity*

What useful things can one say about the countries on the basis of the PPP-derived GDP numbers as they stand? Not much without looking further, because these gross output numbers should be regarded as numerators looking for denominators. Two kinds of denominator candidates immediately come to mind: (i) measures of country need that the gross output must satisfy; and (ii) measures of country effort that went into producing the output. These notions are illustrated in Table 1 for 12 OECD countries included in the 1980 benchmark study.

(i) *Welfare measures*: Column 1 of Table 1 gives each country's GDP per capita, expressed relative to the United States. The concern here is with gross output divided by the number of mouths that must be fed and otherwise maintained out of the gross output. (It would be better to use a needs denominator that takes better account of the relative requirements of different age-sex demographic groups, but such a measure awaits agreement among economic demographers on proper equivalent-adult scales.) Notice that these numbers only bear on judgments about the relative material well-being of countries if it can be assumed that, at least in some gross sense, tastes are the same the world round.

TABLE 1—WELFARE VS. PRODUCTIVITY  
(12 OECD Countries: 1980)

Country	GDP/Pop (U.S. = 1) (1)	GDP/mh (U.S. = 1) (2)	Rank GDP/Pop (3)	Rank GDP/mh (4)
United States	1.00	1.00	1	3
Canada	.99	.93	2	6
Norway	.97	.98	3	4
Germany, F. R.	.87	.90	4	7
France	.86	.93	5	5
Denmark	.86	.84	6	8
Belgium	.85	1.05	7	2
Netherlands	.82	1.07	8	1
United Kingdom	.74	.80	9	9
Japan	.73	.55	10	12
Austria	.71	.73	11	11
Italy	.68	.77	12	10

(ii) *Productivity measures:* Column 2 of Table 1 gives the countries' gross outputs divided by total man-hours of input, also expressed relative to the United States. Of course, alternate divisors can and should be used that take account of other factors of production like human and physical capital. (Incidentally, interesting productivity comparisons have been developed by various researchers for different production-side sub-sectors of many countries on the basis of ICP PPP estimates. The ICP's final-product disaggregate PPPs are not really the right ones to use for quantifying industry-of-origin production, but the right ones are *very* hard to get. Sometimes second-best must suffice.)

Observe that the welfare rankings of the countries as measured by per capita GDP (col. 3) do not match the productivity rankings for GDP per man-hour (col. 4). The rank correlation is .59, which is not very close to unity. Particularly striking are the Belgian and Dutch cases and, for the opposite reason, the Japanese one. If leisure counts, what does this say about the real standing of the United States in welfare terms? As a minimum, the distinction between welfare and productivity becomes blurred. (Caveat: This fairly casual productivity treatment is only meant to be illustrative. A variety of qualifications are called for. To take just one example, consider the possibility of involuntary unemployment. If the leisure differences implicit in the man-hour and population numbers are not a matter of choice, then the GDP/man-hour

figure again might best be interpreted as a productivity measure.) A scatterdiagram of these welfare and productivity estimates shows a rather diffuse cloud ( $\bar{R}^2 = .36$ ) and an overall slope tendency just under one. Textbook hypothesis testing would support a unity slope and zero intercept, but only because of a lack of statistical power. One way of interpreting all of this is to raise the question of whether the ICP's GDP should be regarded as a point on a production-possibility-curve rather than a point on a worldwide indifference map.

### B. Uses of Well-Being Measures

Assume tastes around the world are sufficiently homogeneous to justify thinking GDP per capita is an adequate measure of a country's well-being, so the indifference map notion is appropriate. Then the availability of GDP based upon a common set of prices (the key element of an SRNA) for all political subdivisions is extremely useful in formulating global social policies. Discussions of how international aid should be allocated and how international burdens should be shared are informed in a crucial way by internationally comparable GDP numbers.

A more parochial use of GDP numbers comparable across space and time is in social science research. "Income" is a critical variable in explaining many kinds of human behavior, and in accounting for many kinds of production-side relationships. In addition, national income as a dependent variable gets a great deal of attention in many research areas, even outside the field of economic development.

Consider an investigation of the inter-country world distribution of income based on PPP-based GDP figures (see our article with Kravis, 1984). Similar studies predating the ICP depended on exchange-rate conversions, with the result that the degree of inequality was overstated. If intracountry inequality is ignored (assuming all of each country's GDP is divided equally among its citizens), how unequal is the world's income distribution and how has it changed over the last 35 years? For a world defined by 118 market economies with a total population

TABLE 2—SHARES OF WORLD INCOME: 1950–85<sup>a</sup>

	Income			High
	Low	Oil	Non-Oil	
Number	33	9	54	22
Percent Pop (1985)	44.3	3.9	28.0	23.8
1950	10.8	3.0	13.1	73.1
1960	10.7	3.3	13.9	72.1
1970	9.0	4.3	15.2	71.5
1980	9.3	4.3	19.2	67.2
1985	10.2	3.9	18.8	67.1

<sup>a</sup> For 118 market economies.

covering over 99 percent of the earth outside of the centrally planned bloc, the cross currents displayed in Table 2 were found. The decline in the world share of low-income countries between 1950 and 1980 was reversed after 1980; correspondingly, the growing share of the middle-income countries in the first three decades was arrested; and the share of the industrialized nations stabilized after a long, quite perceptible decline. (Quite significantly, the experiences of countries within both the low- and middle-income groups were far from uniform; inequality increased between 1980 and 1985 *within* each of the groups. All of this is easy to understand if one reviews the large external debt situations in South America and some countries in Africa.

## II. Special Purpose Numbers

The ICP quantities refer to categories of final product rather than to production activity. As noted above, that has not kept some researchers from using ICP PPPs to obtain estimates of real output originating in particular industries. The right industry PPPs are so hard to get that a judicious use of ICP numbers in this way may perhaps be condoned.

We give some examples of useful final-product measures outside the standard national accounts categories that can be developed on a comparable basis across countries from ICP benchmark data. A country's capital stock, estimated from the history of its capital accumulation (and perhaps other

information) can be used for a variety of purposes. However, comparisons of different countries' capital stocks built up out of own-price investment shares (i.e., based on  $I/GDP$  where both  $I$  and  $GDP$  are actual domestic-currency expenditures) will not be right because the prices of construction and producers durables relative to other goods varies so much among countries even at the same income level. This is a general point about investment ratios that applies also to planning models of developing countries and models attempting to explain productivity growth in industrialized economies.

Consider the following investment-ratio example derived from the 1980 benchmark study. In domestic currencies the ratios of investment to GDP for the Philippines, India, Japan, and the United Kingdom were .306, .240, .322, and .169, respectively. If these ratios were construed as indicators of relative quantities, one would be led to believe that the Philippines and Japan were investing about the same proportion of their total output in capital goods, while India and the United Kingdom were at a lower level. However, when the differences in relative prices of capital goods are standardized through the use of the ICP's set of so-called "international prices," the *real* ratios—the *quantities* of capital goods divided by the *quantities* of total output—turn out to be quite different. The real ratios were .148 for the Philippines, .167 for India, .370 for Japan, and .151 for the United Kingdom. The rate of capital formation relative to total output for the Philippines is less than half what was indicated by its own-price ratio, and also half that of Japan. Indeed, it is even less than that of the real ratio for India. Studies attempting to explain productivity changes or aggregate economic growth differences across countries will start out on the wrong foot if they are based on own-price ratios.

Two similar special-purpose examples, involving national defense and service expenditures, involve the same kind of standardization of relative prices. For obvious reasons, there is a great deal of interest in comparisons of defense outlays in different countries. The proportion of a country's output devoted to defense, based on its own

prices, provides a measure of effort that is comparable across countries, but except where price structures are the same, the own-price proportion gives a misleading basis for judging relative *quantities* of military output. The use of judiciously selected ICP PPPs can be of great assistance here. (Heston is doing deflation work of this sort now to get better estimates of defense quantities.) Comparisons are difficult for the latest procurement items, but are quite feasible for personnel, operations, maintenance, and construction, all of which average from 70 to 90 percent of total military expenditures. Taking account of the variability of relative prices across countries in working with military outlays is as important as for other components of GDP. In a sample of 8 OECD countries taking part in a purchasing-power study under the auspices of the United Nations, differences between exchange rate and PPP conversions were substantial. For example, Italy's PPP-derived military outlays were 7.7 percent of that of the United States, while the exchange-rate-derived outlays were 4.6 percent. The equivalent numbers for Australia were 1.5 and 2.6 percent. These differences are good sized even for fairly similar countries, and they are likely to be as large or larger for super-powers and Third World countries.

The place of services in modern economies provides one more example of a special purpose number of great interest. The conventional wisdom has it that rich countries are more service-oriented than poor ones, both with respect to what people in the workforce do and what is consumed. (In fairness, some people view this as more a statement about what happens over time—as nations become richer and whatever else happens over time—than a statement about countries in a cross section.) While the employment part of the conventional wisdom is borne out by a variety of kinds of labor force evidence, the income-elastic part is not consistent with the properly stated facts.

The real share of a country's national output absorbed in the form of services is likely to depart significantly from the ratio of own-price expenditures on services to total own-price expenditures. It is true that when

expenditures are measured in own prices, rich countries take a much larger proportion of their national output in the form of services than do poor countries. However, if allowance is made for the systematic difference in the prices of services relative to goods, then the share of services does *not* go up with country income. A precise statement about the empirical findings here would take some elaboration, but the previous sentence certainly stands as a strong stylized fact. (See Summers, 1985, and references therein.) It takes international comparison data of the ICP type to ferret out this kind of conclusion. It would appear from a preliminary pooled-cross-section analysis of the 1970, 1975, and 1980 benchmark service data that there may indeed be some upward drift not accounted for by changes in prices and income (see our 1987 paper).

### III. Analytical Uses

As already indicated, the most common research use to which ICP international comparisons have been put involve real GDP. Real GDP has found its way into many studies investigating various aspects of energy demand, and has been a critical variable in development studies involving health and education. Section II discussed the use of ICP quantity data in dealing with capital stocks, services, and defense spending. Many of these examples have essentially been special-case demand analyses. In addition, there have been a number of straightforward attempts to estimate, with international data, demand relationships for conventional categories of consumption—food, clothing, shelter, medical care, transportation, communication, recreation, and education. These studies leave unresolved the question of whether at an appropriate level of aggregation one can say with conviction that common tastes around the world dictate that choices differ across countries only because of differences in prices and income. Certainly in simple *ad hoc* regressions at high levels of aggregation, one gets negative price elasticity estimates and reasonable income elasticity estimates. But some taste variables (for example, involving climate and eco-



conomic demography) show up as significant. Not surprising, it is not possible to verify the subtle restrictions of the theory of consumer behavior—Slutsky and all that—in these international data. (See Barten and Summers, 1987).

The above discussion has been concerned with output comparisons. However, a benchmark study is really a pricing exercise because the critical data collected are prices. Many policymakers and researchers have regarded the ICP price data as a useful resource and have used it in a right-hand side way. The analysis of country price structures—here prices are on the left-hand side—is quite interesting, too. Numerical taxonomy clustering exercises have been carried out in which countries are grouped together in accordance with the similarity of their price structures. (See our volume with Kravis, 1982, pp. 104–11.) Similarity was defined in terms of direction cosines of angles between rays out to points representing price vectors in a multidimensional space. The atheoretical clustering algorithm gave entirely plausible results that for the most part are quite in accordance with economic intuition. A blend of income level, trading patterns, and geography seems to determine price structures.

This discussion will end with a comment on a proposition of Paul Samuelson (1974), generally regarded as a plausible theoretically derived fact: *Luxuries are relatively cheap in rich countries and relatively dear in poor countries; and the opposite is the case for necessities.* Is it true? A simple regression can be run to test this with ICP price data on many categories of goods for countries of widely differing incomes. Relative price for a category in a country is placed on the left-hand side of the regression and both country income and category income-elasticity appear on the right, along with an interaction term to allow for the twist implied by the proposition. How does it come out? The proposition is *not* verified, with power aplenty!

#### IV. Conclusion

Apart from the first formal international comparison work in the OEEC in the mid-

1950's, the accumulated experience now touches on almost 70 countries and 3 different years, 1970, 1975, and 1980. (A benchmark study covering only a few countries was done for 1967.) Partial results from a new round ("phase" in the ICP jargon) for the year 1985 have already appeared, but the complete results are still a year away. Two first steps toward a System of REAL National Accounts have been published, and we have written a third. This new so-called Penn World Table (Mark 4) (Mark 2 was never published) provides estimates for 17 variables for the years 1950–85 and 121 market economies (plus just GDP information about 9 centrally planned economies). Whatever the merits of this data set, one thing is clear: a milestone will have been reached and passed in learned journal publication in our March 1988 article. The accompanying new data table will be placed within flaps glued inside the back cover of the *R.I.W.* Electronic publication has come to economics. The data table is on IBM computer-readable floppy diskettes!

#### REFERENCES

- Barten, A. and Summers, R., "An International Demand Model with Varying Tastes," CADE Discussion Paper 87–5, University of Pennsylvania, 1987.
- Kravis, I. B. and Lipsey, R., "National Price Levels and the Prices of Tradables and Nontradables," *American Economic Review Proceedings*, May 1988, 78, 474–78.
- \_\_\_\_\_, Heston, A. and Summers, R., *World Product and Income*, Baltimore: Johns Hopkins University Press, 1982.
- Samuelson, P. A., "Analytical Notes on International Real Income Measures," *Economic Journal*, September 1974, 84, 595–608.
- Summers, R., "Services in the International Economy," in R. P. Inman, ed., *Managing the Service Economy*, New York: Cambridge University Press, 1985, 27–52.
- \_\_\_\_\_, and Heston, A., "Improved International Comparisons of Real Product and Its Composition, 1950–1980," *Review of Income and Wealth*, June 1984, 30, 207–62.

\_\_\_\_\_ and \_\_\_\_\_, "The International Demand For Services," Fishman-Davidson Service Center Discussion Paper 32, University of Pennsylvania, 1987.

\_\_\_\_\_ and \_\_\_\_\_, "A New Set of International Comparisons of Real Product and Prices: Estimates for 130 Countries, 1950-1985," *Review of Income and Wealth*,

March 1988, 34.

\_\_\_\_\_, Kravis, I. B. and Heston, A., "Changes in the World Income Distribution," *Journal of Policy Modeling*, May 1984, 6, 237-69.

United Nations, "World Comparisons of Purchasing Power and Real Product for 1980," ST/ESA/STAT/SER.F/42, 1986.

# National Price Levels and the Prices of Tradables and Nontradables

By IRVING B. KRAVIS AND ROBERT E. LIPSEY\*

The recent debates about the appropriateness of the current U.S. exchange rate and the effects of the decline in the exchange value of the dollar since early 1985 have focused attention on the national price level as an object of economic policy. There seems to be little doubt in the minds of most participants in the debates that the U.S. price level, expressed in any of the world's major currencies, rose greatly from 1980 through early 1985 and since then has fallen sharply. In other words, the purchasing power of a dollar over U.S. goods first declined and then increased relative to the purchasing power of the amount of British, German, and Japanese goods that could be purchased in their own countries with the £, DM, and yen that a dollar exchanged for in currency markets. The Japanese price level, for example, is thus defined as the purchasing power of the yen (the number of yen required to buy in Japan the same goods and services a dollar buys in the United States) divided by the exchange rate (the number of yen required to buy a dollar on the foreign exchange market).

The concepts of purchasing power parity and the "law of one price" are so convenient and so firmly embedded in trade theory and exchange-rate theory that it requires some effort to put them aside to discuss national price levels, but the widely fluctuating exchange rates of the 1970's and 1980's have compelled attention to these issues. Most of that attention has been focused on "real exchange rates" (usually defined as the reciprocal of what we refer to as changes in

national price levels), the relation of exchange rate changes to domestic price changes, but there has been some research centered on explaining the structure of price levels at particular times (Jagdish Bhagwati, 1984; Christopher Clague, 1985, 1986; ourselves, 1983, 1987). One reason for studying the structural or cross-sectional determinants of price levels is that an understanding of the levels from which changes begin is important for explaining the changes.

In this paper we concentrate on the price levels for tradables and nontradables, describing what has happened both in cross sections and over time, and analyze the effects on price levels of changes in income, which we found in our previous work to have been by far the major influence in cross-section analyses of price levels. The role of changes in exchange rates is also briefly described.

## I. The Model

The approach taken here is based on a highly simplified formulation of the factors that determine differences in national price levels across countries and applies it to the comparison among countries of changes over time. In a general-equilibrium setting, the price level is a variable influenced by several long-term or permanent characteristics of an economy. Also, it both influences and is influenced by the various components of monetary and fiscal policy and of international economic policy (choice of exchange-rate regimes and of exchange-rate levels, degree of control over capital movements, extent of autarkic policies).

Across countries, price levels are expected to be positively associated with income because prices of nontradables are higher relative to prices of tradables in rich countries than in poor countries. That might be be-

\*University of Pennsylvania, Philadelphia, PA, 19104 and NBER; Queens College and Graduate Center, CUNY, and NBER, 269 Mercer Street, New York, NY, 10003, respectively. The statistical work for this paper was done by David Robinson and Linda Molinari.

cause productivity differences between rich and poor countries are smaller for nontradables (mainly services) than for tradables, or because nontradables are more labor intensive than tradables and labor is relatively cheap in poor countries (our study, 1983; Bhagwati). The relative price of nontradables would completely determine the price level if prices of tradables in all countries were forced into equality by competition and if the shares of tradables and nontradables in output were identical in all countries. While neither of these conditions holds completely, it is true that prices of tradables, though higher in rich countries because they almost always are sold with an admixture of services, are much more similar among countries than prices of nontradables. Thus the ratio of tradables prices to nontradables prices ( $PTR/PNT$ ) is lower, the higher a country's per capita income. Shares of nontradables in output, the weights of nontradables in total GDP price levels, are either uncorrelated with income levels, and therefore do not affect the income-price level relationship, or are positively correlated with per capita income and therefore reinforce the income-price level relationship (our 1983 study, p. 15).

What implications does the intercountry relationship between income and the  $PTR/PNT$  ratio have for changes in that ratio over time? The intercountry relationship could conceivably reflect either the effects of relative income levels (for example, 90 percent of the U.S. level vs. 45 percent) or the effects of absolute income levels (\$20,000 in international prices vs. \$10,000). If we assume that the relationship observed is with the absolute level of income, we would expect that over time, with rising incomes, the ratios should fall.

What implications, if any, does the intercountry relationship between income and  $PL$  have for changes in  $PL$  over time? Since  $PL$  is defined in relative terms (for example, with U.S. = 100), it is the relative income changes that are pertinent. If a country's per capita income rises relative to the United States, its relative price level should also rise.

There could also be short-run determinants of  $PL$ s and of  $PTR/PNT$ s not oper-

ating through the structural, or long-run relationships we have described. A rise in the exchange value of a country's currency, not offset by changes in domestic currency prices will, by definition, produce an increase in  $PL$ . The assumptions outlined above imply that it should also produce a fall in the  $PTR/PNT$  ratio because prices of tradables are more constrained by international competition than prices of nontradables.

Although real income per capita is the predominant factor influencing the price level, other factors may play some role. In our earlier work we suggested that the degree of openness to trade may also influence price levels and changes in them because a high propensity to trade [ $(X + M)/GDP$ ] pulls a country's prices towards the world average—upward for poor countries and downwards for rich countries. And we expected a high share of tradables in output to reduce a country's price level.<sup>1</sup>

## II. What Do the Cross-Section Data Tell Us about the Relation of Price Levels to Per Capita Income?

The cross-section relations between price levels and income (i.e., real GDP per capita) may be summarized as follows:

1) Price levels tend, as expected, to be lower in poor countries than in rich ones. And also as expected, the differences are greater for nontradables than for tradables.<sup>2</sup> These relationships may be illustrated by the data in Table 1. The simple  $\bar{r}^2$  between the GDP price level ( $PL$ ) and real GDP per capita ( $y$ ) is .52 for the 60 countries.

2) The price level for tradables ( $PTR$ ) rises as income rises, despite the near unanimity found in the literature on real exchange rates that the law of one price prevails for tradables.

3) The relationship between nontradable prices ( $PNT$ ) and real GDP per capita is

<sup>1</sup>For another explanatory framework, including income and several additional factors, but excluding the openness and tradables share variables, see Clague (1985, 1986).

<sup>2</sup>Tradables are defined here as commodity categories in the final expenditures on GDP, less construction.

TABLE 1—INCOME PER CAPITA AND PRICE LEVELS, 1980

	15 Countries with:	
	Lowest Real GDP	Highest Real GDP
Real GDP <sup>a</sup>	\$630	\$9,903
Price Level <sup>b</sup> for		
GDP	69	119
Tradables	80	112
Nontradables	55	129

Source: UN and Commission of the European Communities, 1986.

<sup>a</sup>In 1980 international dollars. For definition of international dollars see Kravis, A. Heston, and R. Summers (1982). Real GDP is per capita.

<sup>b</sup>Base is mean of 60 countries weighted by aggregate real GDP.

stronger than for tradables; the  $\bar{r}^2$  is .66 for nontradables against .31 for tradables. Also, nontradables are much less costly in poor countries and rise more sharply in the progression of countries up the income scale.

4) These relationships produce a negative correlation between the ratio of tradables to nontradables prices and real GDP per capita. This relationship applies even to subsets of countries for which the association between *PL* and *PTR* on the one hand, and real GDP per capita on the other hand is weak or insignificant, notably Africa.

Thus, the cross-section data conform in the main to the model set out above.

Going beyond these simple relationships in which real GDP per capita is the only independent variable, we have recomputed the structural price level equations for *PL*, *PTR*, and *PNT* calculated in previous work for 3-year averages beginning with 1960–62. The calculations performed for this paper are based on the 1980 benchmark data rather than the 1975 survey used previously.

The recalculated equations generally confirm the results in our earlier article (1987). Again the nontradable prices are better explained than tradables. Illustrative of these equations is one for 1982–84:

$$\begin{aligned}
 (1) \quad PNT &= 14.07 + .882y + 7.26OPTI \\
 &\quad (2.26) \quad (7.81) \quad (1.74) \\
 &\quad - .149yOPTI \quad \bar{R}^2 = .836 \\
 &\quad (2.40) \quad RMSE = 10.31
 \end{aligned}$$

where *OPTI* is the ratio (percent) of exports

plus imports of goods and nonfactor services to GDP (*t*-ratios shown in parentheses).

In the entire set of equations, both the openness of the economy and the share of tradables in output played the expected role in structural equations using them, although their effect was not at all as large or as consistent as that for income. Furthermore, while deviations of individual price levels from equality showed little sign of evaporating even after 20 years, the deviations from the levels predicted by our structural equations did tend to diminish substantially. The lower degree of persistence of these deviations suggests that the structural equations come at least a little closer to defining long-run equilibrium than the assumption of identical price levels in all countries.

### III. The Intertemporal Behavior of National Price Levels

If the intertemporal behavior of national price levels could be inferred from the cross-section findings, countries with faster growth in per capita GDP should have larger price level increases and more rapidly declining ratios of tradables price levels to nontradables price levels.

For the period as a whole, price levels rose more in the industrial countries than in the developing countries (an average of 1.74 percent per annum in 11 developed countries vs. 0.48 percent in 33 LDCs). Rates of internal inflation were higher on average in the LDCs but currency depreciation kept the rise in *PL*, *PTR*, and *PNT* in check. For *PL*, for example, a 5.18 percent growth rate in internal (own-currency) prices relative to the United States was offset by a 4.68 percent depreciation, leaving the growth of *PL* at 0.48 percent per annum ( $1.0518 \div 1.0468 = 1.0048$ ). The prices of tradables rose less than the prices of nontradables in developed countries, while the opposite was the case in the LDCs. Thus the *PTR/PNT* ratio of the industrial countries declined relative to that of the LDCs; this was true in both exchange rate subperiods (fixed and floating) and in the period as a whole.

These changes do not necessarily point to the influence of income per capita, even in developed countries where the *PTR/PNT*

TABLE 2—RELATION BETWEEN TRENDS IN INCOME AND TRENDS IN PRICE LEVELS

	Coefficients of Trend in $y^a$		Difference <sup>b</sup>
	PTR (1)	PNT (2)	
All Countries (44)	-.379 (2.1)	-.076 (0.4)	-.303
Industrial Countries (11)	.085 (0.2)	.798 (2.8)	-.713
Developing Countries (33)	-.418 (2.0)	-.201 (1.0)	-.217

Note: Numbers of countries and *t*-statistics are shown in parentheses.

<sup>a</sup>When dependent variable is trend in PTR (col. 1) or PNT (col. 2).

<sup>b</sup>Col. 3 = col. 1 - col. 2.

moved in the expected direction. The same result could have been produced by a more rapid growth of productivity in tradables (goods) production than in nontradables (chiefly services) production.

A direct test of association with income growth can be made by relating trend rates of change in *PL* to trend rates of change in per capita income across countries with the results shown in Table 2. The expected positive relationship is present for the industrial countries, but for the developing countries the relationship is negative.

The expectation regarding the relationship between tradables and nontradables prices fares better. Whatever the direction of the relationship between trends in per capita income and trends in price levels, the coefficient for income is algebraically larger in the equations for tradables. This is also true for the subperiods (before 1971 and afterwards) except in one of the six comparisons that could be made.

The results above for developing countries suggest that changes in exchange rates affect the ratio of tradables to nontradables price levels. We tested that relationship for the floating exchange rate era by relating changes from one 3-year period to another in the tradables/nontradables price ratios to changes in exchange rates. (Three-year averages were used to iron out some of the sharper fluctuations in exchange rates). As predicted, the relationship was consistently negative: a rise in the exchange value of a

country's currency reduced the *PTR/PLN* ratio. The coefficient for the exchange rate change was statistically significant in three out of four periods. For 1978–80 to 1982–84, for example, the coefficient was  $-.249$  (*t*-ratio = 2.69) with an  $\bar{r}^2$  of .18.

The poor predictions of price level change from changes in income alone bring us back to the reason for our interest in the cross-section relationship. That is the expectation that the change in price level in one period will reflect not only the events of that period but how the actual price level compared with that implied by the structural, long-term determinants of price levels at the beginning of the period. That is, if a country's price level was "high" at the beginning of a period, considering its income per capita and other long-term determinants of price levels, we would expect it to fall in subsequent periods, given whatever changes take place in structural variables. We could not find evidence for this over short periods, but it was typically true over periods of 10 years or more. The coefficients of the initial residual in the price equation were usually negative and significant. For example, the equation explaining price level changes between 1960–62 and 1982–84 produced a coefficient of .866 (*t* = 2.97) for the change in per capita income and  $-1.22$  (*t* = 2.87) for the 1960–62 residuals.

Thus, taking fairly long-term movements in price levels, comparing them from one 3-year period to another and using our structural equations to define deviations from equilibrium, we find that changes in income have the expected positive relation to changes in price levels and also that deviations from the "equilibrium" defined by our structural equations have the expected negative relations to subsequent price level movements. The structural equations thus give some meaning beyond purchasing power parity to the idea that a country's price level, and by implication, its exchange rate, can be "too high" or "too low."

#### IV. Conclusion

Across countries, national price levels increase systematically with the level of a country's per capita income, and the ratios

of tradables to nontradables prices decrease. These cross-section relationships carry over to some extent to changes over time. Increases in per capita income are generally associated with increases in price levels in the industrial countries, although the opposite relationship tended to prevail among developing countries. Large increases in income are associated with large declines in the ratio of tradables to nontradables price levels more consistently than with the increases in general price levels, and large increases in the exchange value of the currency are also associated with declines in the price levels for tradables relative to nontradables.

# REFERENCES

- Bhagwati, Jagdish, "Why Are Services Cheaper in Poor Countries?," *Economic Journal*, June 1984, 94, 279-86.
- Clague, Christopher, "A Model of Real National Price Levels," *Southern Economic Journal*, April 1985, 51, 998-1017.
- \_\_\_\_\_, "Determinants of the National Price Level: Some Empirical Results," *Review of Economics and Statistics*, May 1986, 68, 320-23.
- Kravis, Irving B., Heston, Alan and Summers, Robert, *World Product and Income*, Baltimore: Johns Hopkins University Press, 1982.
- \_\_\_\_\_, and Lipsey, Robert E., *Toward an Explanation of National Price Levels*, Princeton Studies in International Finance, No. 52, Princeton University, 1983.
- \_\_\_\_\_, and \_\_\_\_\_, "The Assessment of Nation Price Levels", in Sven W. Arndt and J. David Richardson, eds., *Real-Financial Linkages Among Open Economies*, Cambridge: MIT Press, 1987.
- United Nations and Commission of the European Communities, *World Comparisons of Purchasing Power and Real Product for 1980*, Phase IV of the International Comparison Project (ICP), United Nations and Eurostat, 1986.

# The Sensitivity of International Comparisons of Capital Stock Measures to Different "Real" Exchange Rates

By EDWARD E. LEAMER\*

Gross national investment flows can be discounted and accumulated to form measurements of national capital stocks for cross-national comparisons of capital abundance. The formation of these capital stock figures can depend substantially on the method by which investment measured in different currencies is translated into real investment figures that are comparable across time and across countries. Three different methods for forming time-series of real investment figures are considered here. One of these methods translates foreign currencies into dollars in each year using the current exchange rate and then divides by the U.S. deflator for gross investment to form a series on real investment. The second method accumulates real investment in the local currency deflated by the home country price deflator for gross national investment and then translates into dollars using the base year (1966) exchange rate. The third measure of capital substitutes the purchasing power parity rates from Robert Summers and Alan Heston (1984) in place of observed exchange rates. These three capital stock comparisons are contrasted and used here in a latent variable model of trade. The conclusion from this data analysis is that although the measures of capital stock can differ substantially for some countries, these differences matter very little when capital stock is treated as one variable in explaining the composition of trade. The data do suggest some slight preference for the third measure of capital based on the purchasing power parity exchange rates.

## I. Formula for Capital Stock Comparisons

The data consist of time-series on the following numbers:  $I_t$  = gross domestic investment in home country currency referred to here as "pounds";  $P_t$  = home country investment price index;  $P_t(\$)$  = U.S. investment price index;  $S_t$  = exchange rate, pounds, per U.S. dollar;  $PPP_t$  = purchasing power parity exchange rate, pounds per U.S. dollar = the pound price of a commodity bundle divided by the dollar price of the same bundle.

Transformations of the basic data into a common currency are required to compare investment figures across countries. One possibility is first to control for changes in the pound price index by translating the year  $t$  investment into year  $b$  (base year) pounds using the pound investment price deflator, and then to divide by the base-year exchange rate to form a figure for year  $t$  investment in year  $b$  dollars:

$$(1) \quad R_t(1) = I_t(P_b/P_t)/S_b,$$

where  $R$  is a mnemonic for real investment. Alternatively, the year  $t$  investment can be translated into year  $t$  dollars, and then the dollar investment deflator can be used to translate into year  $b$  dollars:

$$(2) \quad R_t(2) = (I_t/S_t)(P_b(\$)/P_t(\$)).$$

The third measure of real investment uses the Summers-Heston series on  $PPP$  exchange rates to convert to U.S. dollars year by year:

$$(3) \quad R_t(3) = (I_t/PPP_t)(P_b(\$)/P_t(\$)).$$

The investment series  $R(1)$  and  $R(3)$  can be thought to correct the investment series

\*Department of Economics, UCLA, Los Angeles, CA 90024.



$R(2)$  for misalignment of the exchange rate. The first investment series is equal to the second divided by an index of the real exchange index:  $R_i(1) = R_i(2)/(r_i/r_b)$ , where the "real exchange rate" is  $r_i = S_i P_i(\$)/P_i$ . These real investment series would be identical if the real exchange rate were constant,  $r_i = r_b$ , that is, if variation in the relative price of investment goods were exactly offset by variation in the exchange rate. The third real investment series is a corrected version of the second:  $R_i(3) = R_i(2)/P_{it}$ , where  $P_{it} = PPP_i/S_i$  is the ratio of the purchasing power parity rate to the exchange rate. (This is what Heston and Summers refer to as the price index for gross domestic investment.) The correction factor  $1/P_{it}$ , which is defined to be equal to one for the United States, is generally greater than one for the developing countries, and is often slightly lower than one for the developed countries.

This paper is concerned with the sensitivity of capital stock estimates to choice of real exchange rates. Other issues that are important for cross-country comparisons of capital stocks are the choice of survival profile and the level of disaggregation. The survival profile can be important because the intertemporal investment patterns are substantially different for many countries. The 15-year lifetime that is used here emphasizes relatively recent investments and overstates the capital stocks in countries with high current savings rates. Disaggregation of the investment flows can also be important for two reasons. First, if survival profiles differ depending on the form of the investment and if the composition of investments differs substantially across countries, then the aggregate survival profile will differ substantially across countries. Secondly, aggregation of capital into a single measure is implicitly based either on the assumption that capital is fungible or that the relative prices of the components of the capital stocks do not change over time. Disaggregation is therefore called for when survival profiles differ for different components of investment and when relative price variability is substantial.

Here, a particular survival profile is selected and three different capital stock measures are computed using the three dif-

ferent real investment series defined above. The three real investment series are accumulated assuming exponential survival profiles to form three real capital stock series in year  $b$  dollars:

$$(4) \quad K_i^b(j) = \sum_{j=t-\theta}^t (1-\delta)^{t-j} R_i(j),$$

$j=1,2,3,$

where  $\delta$  is the rate of depreciation and  $\theta$  is the length of asset life. Here the asset life is assumed to be 15 years and the rate of depreciation is assumed to be 13.3 percent, a number commensurate with the double declining balance method with a 15-year life.

The capital stock figure  $K(1)$  uses the base-year exchange rate to translate all the investment flows into dollars. If in this base year the dollar were under- or overvalued relative to the pound, the British capital stock figure would be correspondingly over- or undervalued for all time periods. The figure  $K(2)$  has an analogous sensitivity to the base-year dollar price index,  $P_i(\$)$ , which is however less likely to be as sensitive to "monetary disturbances" as the exchange rate. But in effect  $K(2)$  treats the investment bundles in the United States and other countries as if they were the same, which may be subject to substantial doubt especially for developing countries.

These problems of cross-country comparisons of investment rates are implicitly dealt with when the prices corrected for purchasing power parity are selected. If investment goods were priced the same throughout the world and if we had data for each country on the quantities of each type of investment, then it would be a straightforward exercise in index number construction to form indices of real investment for cross-country comparisons. The work on purchasing-power-parity price indices of Heston-Summers implicitly accepts this as a premise though it confronts measurement difficulties. It is accordingly unclear at the outset that the measure  $K(3)$  is actually superior to the other two measures of capital.

## II. Comparisons of the Three Measurements of Capital Stocks

### A. Observations Regarding the Time-Series Behavior of the Investment and Capital Stock Figures

My estimates of capital stock for as many as 110 countries begin in 1975, 15 years after the first period in which the investment data are reasonably complete. These capital stock figures are computed through 1980. The time-series correlations between  $K(1)$ ,  $K(2)$ , and  $K(3)$  over this 6-year period are high for most of the 110 countries. The average correlations are  $(\rho_{12}, \rho_{13}, \rho_{23}) = (.87, .86, .85)$ . These time-series correlations are not high for all countries; in fact they are negative for several "peripheral" countries for which data are suspect and/or there have been large changes in the real exchange rates: Uganda, Upper Volta, Angola, Zaire, Ethiopia and New Zealand, to name a few. For these countries and several others it will be important to decide which is the better measure of capital stock.

### B. Observations Regarding Cross-Section Correlations

For cross-country comparisons it is necessary to eliminate the effects of country size, which can be done by dividing the capital stocks by the World Bank series for total population. An alternative divisor would be GNP, but GNP is subject to the same measurement problem that is the concern of this paper. The data for 1980 displayed in Table 1 are typical. The capital per man based on the measure  $K(2)$  has a larger mean, a larger standard deviation and a larger range than the other two measures. This is to be expected because the other two measures can be thought to correct  $K(2)$  for erratic swings of the exchange rates.

The 1975 correlations of the three capital per GNP figures are  $(\rho_{12}, \rho_{13}, \rho_{23}) = (.93, .70, .66)$ . The last two of these numbers are about the same in later years but the first slowly deteriorates to .78 in 1980 as the period of exchange rate variability begins to affect the measure of capital stocks on GNP.

TABLE 1—1980 STATISTICS: CAPITAL/POPULATION (105 countries)

	Mean	Standard Deviation	Minimum Value	Maximum Value
$K(1)/POP$	1325	1712	33	6757
$K(2)/POP$	1626	2171	34	8207
$K(3)/POP$	1594	1758	59	5887

### C. Identifying Outliers

Potentially unusual observations are identified first by regressing capital per man figures in 1980 for each of the three ways of measuring capital against each other. The regression of  $K(1)/POP$  on  $K(3)/POP$  has an  $R^2$  of .96. The largest Studentized residuals of this regression are: Iceland (6.4), Gabon (-3.7), Sweden (3.2), Portugal (-2.8), and Trinidad-Tobago (-2.1). Thus compared to the capital stock measure using purchasing power exchange rates, the first measure of capital per man ( $K(1)/POP$ ) is significantly higher for Iceland, Sweden, and significantly lower for Gabon, Portugal, and Trinidad-Tobago. The regression of  $K(1)/POP$  on  $K(3)/POP$  has an  $R^2$  of .98. The largest Studentized residuals of the regression of  $K(2)/POP$  on  $K(3)/POP$  are Switzerland (4.6), Norway (3.6), Portugal (-2.9), Iceland (2.4), and Sweden (2.1). These later results are suggestive of the relative downward adjustment of the capital stock figures for the developed countries when the purchasing power parity exchange rates are used.

Potentially unusual observations are also identified by regressing per capita GNP on capital per man. If capital and labor were the only factors of production, and if factor prices were equalized internationally, then  $GNP/LABOR = w + r \text{ CAPITAL}/LABOR$ , where  $w$  is the wage rate and  $r$  is the rental rate of capital in suitable units. An outlying country in the bivariate scatter of  $GNP/LABOR$  and  $CAPITAL/LABOR$  can therefore be explained by reference to either: 1) other factors which contribute substantially to GNP, 2) differences in factor prices, or 3) mismeasurement of CAPITAL, GNP, or LABOR.

TABLE 2—LARGEST STUDENTIZED RESIDUALS, 1980

K(1)		K(2)		K(3)	
Kuwait	7.46	Kuwait	13.1	Luxembourg	4.13
Iceland	-3.70	Saudi Arabia	2.99	Gabon	-4.11
Canada	-3.28	Gabon	-2.45	Japan	-3.37
Luxembourg	2.85	Japan	-2.17	Switzerland	3.37
Libya	2.8	Norway	-1.96	Sweden	2.72

Note: Regression of  $GNP/POP$  on  $K(j)/POP$ .

Regression of per capita  $GNP$  on capital per man in 1980 yields coefficients for the three capital series of  $(w, r) = (22.8, 2.58)$ ,  $(126.9, 2.05)$ ,  $(-432, 2.33)$ . The  $R^2$ s are about the same for  $K(1)$  and  $K(2)$ , and somewhat higher for  $K(3)$ . The negative estimates of the constant when  $K(3)/POP$  is the explanatory variable casts doubt on this measure of capital, but keep in mind that  $GNP$  uses the the current exchange rate, not the purchasing power parity rate.

The unusual observations are reported in Table 2, where Studentized residuals are used as measures of unusualness. The residuals are biggest for several of the oil producing countries. This suggests not mismeasurement but rather the omitted factor: oil reserves. Incidentally, the country coverage is somewhat greater for  $K(1)$  and  $K(2)$ ;  $K(3)$  in particular excludes the oil-producing countries. Also the  $GNP$  figure uses current exchange rates, not the purchasing power parity rates.

### III. Combining the Capital Stock Measures using a Latent Variable Model

Examination of details of the three capital stock series can be helpful but not conclusive in choosing among the alternative measurements. Further evidence about the accuracy of the three series can be obtained when they are put to use to explain some other phenomenon. Here I report a Heckscher-Ohlin model that explains the commodity structure of international trade in terms of the availability of resources such as labor, land, and capital. First, the three capital stock measures are used one at a time to determine if one is vastly superior in explaining trade. Secondly, capital is treated as an unobservable, or "latent" variable and the overall

model is estimated with LISREL. This latent variable model implies an overall index of capital that combines optimally the three capital stock measures series and, if desired, also the data on trade.

#### A. A Theoretical Model

A heteroscedastic Heckscher-Ohlin-Vanek trade model (estimated in my 1984 book) is  $T_{ij} = \beta_j' X_i + \theta_j K_i + GNP_i \epsilon_{ij}$  where  $j$  is the commodity subscript,  $i$  is the country subscript,  $X$  is a vector representing the supply of  $m$  factors other than capital,  $K$  is the supply of capital,  $GNP$  is gross national product, and  $\beta$ ,  $\theta$ , and  $\epsilon$  are unobservables, the latter assumed to be normally distributed with mean zero and variance  $\sigma_j^2$ . This model is intended to explain the cross-country variability in the commodity composition of trade at one point in time.

#### B. Regression Estimates

First, the linear trade system explaining the same ten aggregate commodities (defined in my book) was estimated using as explanatory variables, separately each of the three capital stock figures, the rest of the variables (also defined in my book) plus the level of the trade balance. If an unweighted model is used,  $K(3)$  gives the highest  $R^2$  in eight of the ten regressions, though the differences are not substantial. But if the data are weighted by  $GNP$  or by population, the results are mixed with each of the measures of capital stock performing about the same. The conclusion that seems warranted is that more careful consideration of the method of translating investment into a common currency does not produce dramatic improvements in our understanding of the determinants of the commodity structure of trade.

#### C. LISREL Estimation

Next the true capital stock is treated as a latent variable and the LISREL program is used to estimate the system. There are three direct measurements of investment flows, differing in the way that local currencies are translated into dollars. A natural model for

this measurement process is  $M/GNP = \alpha + \gamma K/GNP + e$ , where  $M$  is a three-by-one vector of measurements of the capital stock and  $K$  is the true capital stock.

Estimation of this latent variable model was done for data weighted by  $GNP$  and by population. The results do not favor dramatically any of the measures of capital, but the factor loadings are frequently highest for  $K(3)$ . (Details are available on request.)

## REFERENCES

- Leamer, Edward E., *Sources of International Comparative Advantage: Theory and Evidence*, Cambridge MIT Press, 1984.
- Summers, Robert and Heston, Alan, "Improved International Comparisons of Real Product and Its Composition: 1950-1980," *Review of Income and Wealth*, June 1984, 30, 207-62.

AMERICAN ECONOMIC ASSOCIATION

---

PROCEEDINGS  
OF THE  
ONE-HUNDREDTH  
ANNUAL  
MEETING

CHICAGO, ILLINOIS  
DECEMBER 28–30, 1987

## The John Bates Clark Award

*Citation on the Occasion of the Presentation  
of the Medal to*

SANFORD GROSSMAN

*December 29, 1987*

Sanford Grossman's research has had a profound influence on the pure theory of markets as well as on financial economics, aspects of labor economics, and industrial organization. His work is unified by a concern with the economics of information and the efficiency tradeoffs that are attributable to conditions of uncertainty and information asymmetry. His early contributions demonstrated the fundamental tension between the social benefits of displaying the value of information through market prices and the private costs of traders in collecting costly information. An examination of labor market contracting revealed that unemployment may be an unavoidable result of the impossibility of conditioning wages on state realizations. His research (much of it in collaboration with Oliver Hart) has brought new insights into the relations between corporate financial structure and managerial incentives and the market for corporate control (takeover). More recent work on incentives, information asymmetries, and non-contractibility has opened up an entirely new area of research on incomplete contracting. Grossman's research is everywhere marked by keen economic intuition and the use and development of rigorous modeling techniques to study the leading information economics/complex contracting problems of the day.

# Minutes of the Annual Meeting

## Chicago, Illinois

### December 29, 1987

The one-hundredth Annual Meeting of the American Economic Association was called to order by President Gary Becker at 6:15 P.M., December 29, 1987 in the Grand Ballroom of the Chicago Hyatt Regency Hotel. The minutes of the meeting of December 29, 1986, were approved as published in the *American Economic Review, Papers and Proceedings* (May 1987, p. 361).

The Secretary (C. Elton Hinshaw), Treasurer (Rendigs Fels), Editor of the *American Economic Review* (Orley Ashenfelter), Editor of the *Journal of Economic Literature* (John Pencavel), Editor of the *Journal of Economic Perspectives* (Joseph Stiglitz) and the Director of *Job Openings for Economics* (Hinshaw) discussed their written reports published elsewhere in this issue. These had been distributed to members prior to the meeting.

Becker announced the retirement of Rendigs Fels as Treasurer of the Association and read the following resolution passed by the Executive Committee:

Whereas

Rendigs Fels made his first report as Secretary and Treasurer to the AEA Executive Committee on December 27, 1970. And whereas today, he makes his

last report as Treasurer of the Association. And whereas seventeen Executive Committees and Presidents have benefited from his guidance, advice, and good judgment. For 17 years he has carried out his duties ably and responsibly. We have been privileged to have had his company and counsel for so long. He has served the Association with great distinction and can depart in the sure knowledge of a job well done. He has dealt with a disparate group of economists with courtesy, patience, and good humor.

Therefore, be it resolved, that the Executive Committee, on behalf of the Association, express its gratitude and appreciation for Ren with a standing round of applause.

The meeting responded with another round of standing applause.

Becker then introduced Robert Eisner (who happened to be out of the room at the moment) as the 1988 President of the Association. There being no further business, the meeting was adjourned.

Respectfully submitted,  
C. ELTON HINSHAW, *Secretary*

## Minutes of the Executive Committee Meetings

**Minutes of the Meeting of the Executive Committee in New York, New York, March 20, 1987.**

The first meeting of the 1987 Executive Committee was called to order at 10:05 A.M. on March 20, 1987 in the Miller Room of the New York Marriott Marquis Hotel, New York, New York. Members present were Gary S. Becker (presiding), Orley Ashenfelter, Robert Barro, Alan Blinder, Robert Eisner, Rendigs Fels, Ann Friedlaender, Elton Hinshaw, Charles Kindleberger, Robert E. Lucas, Jr., John Pencavel, Sherwin Rosen, and Judith Thornton. Present for parts of the meeting were members of the Nominating Committee (Marjorie Honig, Stephen Magee, Frederic L. Pryor, Michael Rothschild, Richard Schmalensee, Charles Schultze, and Frank Stafford) and the Committee on Honors and Awards (James J. Heckman, Dale W. Jorgenson, Richard R. Nelson, John B. Taylor, William Vickrey, and Oliver Williamson). Also present were Leo Raskind, Counsel, and Carl Shapiro, Co-Editor of the *Journal of Economic Perspectives*.

Becker opened the meeting by informing the group that the National Science Foundation, the Sloan Foundation, and the Ford Foundation had requested the Association to undertake a study of graduate education. These foundations had expressed a concern that current graduate training emphasized abstract, mathematical modeling at the expense of applied work with an historical and institutional perspective. Becker indicated that he had some sympathy for that point of view but has reservations—the market doesn't seem to reflect such a disenchantment; there is little to no hard evidence to support that view; the issue is not new (each new generation of economists seems too abstract and mathematical to the older one); and there is so wide a divergence of opinion in the profession that a consensus could probably not emerge from a study. Nevertheless, the perception exists within the foundations, which are among the major sources of funding of economic research,

that the concern is justified. They are willing to fund a study of graduate training.

It was decided to appoint a committee to assay graduate training. The committee would gather information on curricula and students and faculty perceptions about the content and direction of graduate education. The emphasis of the committee's study was to be on "What is." Serious doubts were expressed that a consensus could (or should) be reached on "What ought to be."

*Minutes.* The minutes of the meeting of December 27, 1986, were approved as written and circulated prior to the meeting.

*Report of the Secretary* (Hinshaw). The Secretary reminded the Committee that the schedule for future annual meetings of the association is Chicago in 1987, New York in 1988, and Atlanta in 1989. San Francisco and Washington, D.C. are the leading prospects for the 1990 site.

Registration for the 1986 New Orleans meeting totaled 6,135, a record for a non-East Coast site. Thirty-eight other associations, societies, and organizations met with us, 357 scholarly sessions were held, and 111 other events (parties, committee meetings, etc.) were scheduled. In New York (1985), registration was 7,349; 38 other groups participated; 355 scholarly sessions were held; and 94 other events scheduled.

He reported that he would submit an informational ballot (nonbinding) to the members concerning preferred dates for the annual meeting. The last such ballot (1978) indicated that, among six possibilities, the traditional Christmas-New Year period ranked first followed closely by the weekend immediately after New Year's Day. This ballot would restrict the choice to one of these two.

*Report of the Editor of the American Economic Review* (Ashenfelter). It was VOTED to approve Ashenfelter's recommendation for the appointment of Maurice Obstfeld to the Board of Editors. He announced the resignation of John Riley as one of the Co-Editors. The Committee noted with thanks



Riley's valuable contributions to the journal, first as Associate Editor with Robert Clower and then as Co-Editor with Ashenfelter. Again, acting on Ashenfelter's recommendation, it was VOTED to approve Hal Varian as a new Co-Editor to replace Riley.

Ashenfelter then presented a research proposal (a copy of which is available from the Secretary) concerning a double-blind refereeing experiment for the *AER*. The study would be conducted by Rebecca Blank of Princeton University; it would provide experimental data on the effect of single- vs. double-blind refereeing and would focus on the impact of gender and institutional affiliation on referee evaluations. It was VOTED to approve the experiment even though many members of the Committee indicated that they thought the Association should adopt double-blind refereeing regardless of the outcome of the study.

Ashenfelter announced the retirement of Wilma St. John as Production Editor. It was VOTED to express the Association's deep appreciation for her excellent work and unusual dedication to the welfare of the journal. It was noted that she had trained three *AER* editors, George Borts, Robert Clower, and Ashenfelter.

*Report of the Editor of the Journal of Economic Literature* (Pencavel). Pencavel again raised the issue of the increasing cost of the journal and the need for a thorough review of the bibliographic classification scheme. Continued expansion in the number of journals in the discipline has increased the costs of bibliographic work and printing. He is seeking ways to restrain costs without seriously affecting the usefulness of the journal. Of course, another option is to allow expansion and increase funding.

*Report of the Editor of the Journal of Economic Perspectives* (Shapiro). Shapiro reported in the absence of Joseph Stiglitz, the Editor. The first issue is expected to be published this summer. It will contain papers from two symposia—one on tax reform, the other on arbitrage. Special features include "Economic Puzzles" and "Recommendations for Further Reading." The journal is slowly building an inventory of accepted articles. The editors now have enough material for the next four issues. Shapiro

noted that the journal's wide circulation had been important for acquiring papers.

*1987 Program Committee* (Eisner). President-elect Eisner reported that he had received over 300 session proposals for the program, the basic theme of which is "The Challenge of Full Employment." He has organized or agreed to cosponsor about 115 sessions. He was pleased to announce that Alan Blinder had agreed to give the Richard T. Ely Lecture.

*Committee on Honors and Awards* (Williamson). Williamson, Chair of the Committee, reviewed the winnowing process used to select candidates for the John Bates Clark Medal. The Electoral College, consisting of the Committee on Honors and Awards and the Executive Committee, VOTED to award the 1987 Clark Medal to Sanford Grossman, Princeton University.

*Nominating Committee* (Schultze). The Electoral College, consisting of the Nominating and Executive Committees meeting together, chose Joseph Pechman as the nominee for President-elect and Arthur Goldberger and Thomas Schelling as Distinguished Fellows. Schultze reported the Committee's choices for other offices: for Vice-President (two to be elected), Martin Feldstein, F. M. Scherer, Roy Radner, and Michael Spence; for members of the Executive Committee (two to be chosen), George Akerlof, Edward Gramlich, William Brock, and Isabel Sawhill.

*Other Business* (Becker). It was VOTED to renew the terms of both Ashenfelter, Editor of the *AER*, and Pencavel, Editor of the *JEL*, for another three years. Their terms of office will now expire December 31, 1991.

*Report of the Treasurer* (Fels). Fels distributed to the Executive Committee the audited financial statements for 1985 and 1986. The "Statements of Revenues and Expenses" show an operating deficit of \$129 thousand for 1986, down from \$164 thousand the previous year. After taking into account investment gains of \$248 thousand in 1986, there was a surplus of \$119 thousand.

As previously agreed, the method of calculating investment gains has been changed. The formula used for 1986 (which will be used for future years) is 5 percent of cash and investments at the beginning of the year.

As a result, the figures for investment gains for 1985 and 1986 in the income statement are not comparable. The 1985 figure was calculated by a formula that included dividends, interest, and inflation-adjusted capital gains (whether realized or not) with capital gains on equities recognized in equal annual installments over three years. Under the old formula, investment gains for 1986 would have been \$545 thousand compared to \$449 thousand in 1985, and the 1986 surplus would have been \$434 thousand in 1986 compared to \$285 thousand the previous year.

The ratio of net worth as of December 31, 1986, to budgeted expenditures for 1987 was 1.64. A ratio of 1.0 is considered ample for safety. The net worth at the end of 1986 was \$3,982 thousand. The budget adopted by the Executive Committee at its meeting on December 27, 1986, calls for expenditures of \$2,426 thousand this year. Thus, the excess over safety requirements is on the order of a million and a half, ample to finance the new journal (the *Journal of Economic Perspectives*), for which the Executive Committee has provided half a million for startup costs.

The budget for 1987 adopted by the Executive Committee on December 27, 1986, showed an operating loss of \$413 thousand, investment income of \$182 thousand, and an overall deficit of \$231 thousand. Now that audited figures for the end of 1986 are available, the investment income for the year is known, viz., \$271 thousand, reducing the projected deficit to \$412 thousand. The deficit can be expected to rise each year unless action is taken to increase revenues.

Acting upon the Budget Committee's recommendation, it was VOTED to increase subscription rates from \$110 to \$125 and keep membership dues unchanged. It was also VOTED to establish a committee to investigate the feasibility and wisdom of transferring management of the Association's portfolio from the investment counselor to the Treasurer. It was understood that, if the transfer was made, the Treasurer would be restricted to "Indexed Funds"; selection of particular stocks and bonds would be eliminated. The Finance Committee would continue its supervisory role.

There being no further business, the meeting adjourned at 2:50 P.M.

#### **Minutes of the Meeting of the Executive Committee in Chicago, Illinois, December 27, 1987:**

The second meeting of the 1987 Executive Committee was called to order at 10:10 A.M. on December 27, 1987, in the Horner Room of the Chicago Hyatt Regency Hotel. Members present were Gary S. Becker (President), Orley Ashenfelter, Robert J. Barro, Alan S. Blinder, Robert Eisner, Rendigs Fels, C. Elton Hinshaw, Charles P. Kindleberger, Robert E. Lucas, Jr., Daniel McFadden, John Pencavel, Alice Rivlin, Sherwin Rosen, Thomas J. Sargent, Joseph Stiglitz, and Judith Thornton. Newly elected members of the 1988 Executive Committee present as guests were Martin Feldstein, Joseph A. Pechman (President-elect) and F. M. Scherer. Present for part of the meeting were Nancy Gordon, Michael McCarthy, and Ronald Oaxaca. The Association's legal counsel, Leo Raskind, also attended.

Becker opened the meeting by welcoming the new members of the 1988 Executive Committee (Pechman, Feldstein, and Scherer) and thanking those whose terms had expired (Kindleberger, Lucas, Blinder, McFadden, Fels) for their fine service to the Association. He then reported on the current standing of the prospective study of graduate training in economics discussed at the March meeting. He reported to the interested foundations that the Association was willing to undertake a descriptive study of the state of graduate training, but would need financial support. If they were interested in funding such a study, they should let him know. He has heard nothing since. Taking this as a lack of interest on their part, he has not pursued the idea. He has since noted that the NSF has instituted a new program for predoctoral and postdoctoral support for empirical economics, perhaps in partial response to their interest in the type of graduate training taking place. He then reviewed the agenda. After an extended discussion of the issues involved in a study of graduate training, it was decided to check again with the foundation to see to what

extent they might be interested in funding a factual study as proposed by the Executive Committee.

*Minutes.* The minutes of the meeting of March 20, 1987, were approved as written and circulated prior to the meeting.

*Report of the Secretary* (Hinshaw). The Secretary reminded the Committee that the meetings are in New York in 1988 and Atlanta in 1989. Acting on the recommendation of the Secretary, the Executive Committee VOTED to approve Washington as the site for 1990 and authorized him to enter into contracts with the hotels.

The results of the preferential ballot concerning meeting time and sites were reported.

Table 1—Dates of Meeting

Options	Votes	Rank
Strongly prefer		
December 27–30	1,028	2
Strongly prefer 1st		
weekend after January 1	2,103	1
Prefer December 27–30	611	5

Table 2—Preferred Sites

City	Votes	Rank
Area I		
Baltimore	876	5
Boston	2,804	2
New York	2,546	3
Philadelphia	1,088	4
Washington, D.C.	3,236	1
Area II		
Atlanta	2,080	3
Chicago	2,575	2
Cincinnati	861	5
New Orleans	3,128	1
San Antonio	1,671	4
Area III		
Las Vegas	845	5
Los Angeles	1,240	4
San Diego	2,115	3
San Francisco	3,944	1
Seattle	2,296	2

In keeping with the members' preferences, it was VOTED to instruct the Secretary to explore the feasibility of moving the 1991

meeting to the first weekend in January (1992) after January 1. The Executive Committee would meet on a Thursday and the meetings would take place on Friday, Saturday, and Sunday.

He reported that the postal authorities have ruled that the Association is not complying with all second-class mailing regulations. One of the regulations requires that the lowest membership rate be not less than 50 percent of the subscription price. Currently the lowest membership rate is \$38.50 and the subscription rate is \$125. The Nashville post office is considering the issue.

The next meeting of the Executive Committee will be on Friday, March 18, 1988, in New York at the New York Hilton.

*Report of the Editor of the American Economic Review* (Ashenfelter). Ashenfelter reviewed his written report (see elsewhere in this issue). Acting on his recommendation, it was VOTED to appoint the persons listed below to the AER Board of Editors: (for three-year terms) James Anderson, Kenneth Judd, Barbara Spencer, David Sappington, Robert S. Smith, Claudia Goldin, and John G. Riley; (for one-year terms) Leslie Young, George E. Johnson, and John Kennan. He also gave a status report on the double-blind refereeing experiment being conducted by Rebecca Blank. As papers are received they are divided into two groups—"blind" and "not blind." Information on authors is removed from the blind group prior to their being sent to referees and a brief questionnaire is included for the referee asking, among other things, whether or not the referee knows (or can determine) the author. Since the beginning of the experiment, 529 papers have been received. Few of these papers have received a final decision (73 have been rejected and 1 accepted). Clearly it is too soon to provide results. Another status report will be given at the March meeting.

*Report of the Editor of the Journal of Economic Literature* (Pencavel). Pencavel briefly reviewed his written report (see elsewhere in this issue). Acting on his recommendation, it was VOTED to approve the appointment of Harvey Rosen and Mark Rosenzweig to the JEL's Board of Editors. He also pro-

posed that marketing of the *Index of Economic Articles* be undertaken by the Pittsburgh office under the direction of Drucilla Ekwurzel. Richard D. Irwin has terminated its contract with the Association to distribute the *Index*. Investigation of several alternatives led him to believe that the Pittsburgh office could best handle the sales and order-handling process. It was VOTED to approve his recommendation.

One of the possibilities being considered for restraining the increase in net cost of the journal was the publication of the abstracts as a separate "journal" to which persons would have to subscribe. After an extended discussion of the merits of an "abstracts journal" and variations on same, it was decided to do nothing for the immediate future.

*Report of the Editor of the Journal of Economic Perspectives* (Stiglitz). Stiglitz briefly reviewed his written report (see elsewhere in this issue). The major event, of course, was the publication of the first two issues. The two totaled 400 pages in length, included 20 full-length articles, symposia on tax reform, merger policy, and arbitrage, and contained three short, regular features—Recommendations for Further Reading, Puzzles, and Anomalies. Stiglitz also raised two policy issues: the handling of unsolicited manuscripts and the possibility of publishing the annual Richard T. Ely Lecture and the Presidential Address in the *JEP*. It was confirmed that the *JEP* should continue to place major emphasis on solicited manuscripts as originally intended and that the *AER* should be considered the appropriate journal for unsolicited articles that require refereeing. Unsolicited manuscripts and proposals for articles and symposia that find their way to the *JEP* should be considered carefully but the journal should not begin to evolve into a standard, refereed journal. It was decided to postpone a definite decision about where to publish the Ely Lecture and the Presidential Address until a committee has evaluated the *JEP* and considered a means of establishing permanent financing for it. The sense of the meeting was that both should probably remain in the *AER*. It was VOTED to approve

the appointment of Dwight Jaffee, John E. Roemer and Kenneth S. Rogoff to the *JEP*'s Board of Editors.

*Report of the Committee on the Status of Minorities in the Economics Profession* (Oaxaca). Oaxaca first reviewed the two AEA Minority Fellowship programs, the Rockefeller Fellowships and the Federal Reserve Bank (FED) Fellowships. The Rockefeller program funds minority students during the first two years of graduate study; the FED program provides two years of dissertation support. There are currently fellows studying at nine different universities. The Rockefeller funding will end December 31, 1987. Once funding for the summer program is found, the Committee will seek new funds to replace the lost Rockefeller monies. The summer program is currently funded by the Sloan Foundation, but the grant expires in 1989. Seeking funds for the summer program is the first priority.

He further noted that, based on National Research Council data for the 1976 to 1986 period, the number of Ph.D.s conferred on minorities had increased both absolutely and relatively, from 1.5 percent of the total in 1976 to 3 percent in 1986. He believes the AEA program has helped sustain and support this trend.

McCarthy reviewed his lengthy written report on the 1987 Summer Program at Temple University. (A copy is available from the Office of the Secretary.) He noted that a continuing problem for the program is the wide range of math training among the participants and sought opinions about how to deal with it. It was agreed that adequate training in math was absolutely necessary for prospective graduate students. Two suggestions were made: (1) make calculus a prerequisite for entry into the summer program, and (2) institute a 3-week intensive "remedial" math program prior to the main program.

*Report of the Director of Job Openings for Economists* (Hinshaw). Hinshaw referred the Committee to his written report (see elsewhere in this issue) and simply noted that the number of new jobs listed reached an all-time high of 1,924 this year.

*Report of the Committee on Foreign Honorary Members.* Richard Caves, chair of the committee, submitted a written report ranking various candidates for election to honorary membership in the Association. It was VOTED to elect Alexander Cairncross, W. M. Gorman, Ryutaro Komiya, and Mario Simonsen as foreign honorary members.

It was also VOTED to transfer the responsibility for nominating persons for this distinction to the Committee on Honors and Awards and discharge the existing committee. Since the Committee on Honors and Awards meets regularly to nominate for the Clark Medal, it was believed that this action would allow for a more systematic review of foreign economists and simplify the Association's committee structure. The committee would be expanded to include persons with substantial familiarity with foreign economists. It was understood that future nominations would be made every other year as with the Clark Medal.

*The 1988 Program* (Pechman). Pechman announced that the theme for the 1988 program will be "Economics for the New President." He has begun contacting persons to organize sessions emphasizing current economic issues.

*Other Business* (Hinshaw). Hinshaw asked for a discussion of two letters he had received from members seeking special dispensation regarding membership dues. One letter requested a free membership for a third party and the other asked for a lifetime membership at a very low rate. The current policy is that "special" rates are not available. The Secretary noted that he received such requests with some regularity and wanted to make sure that the existing policy has the approval of the Executive Committee. These two letters were samples of the general type. It was VOTED to continue the existing policy.

*Resolution* (Eisner). Prior to the report of the Treasurer, Eisner offered the following resolution:

Whereas

    Rendigs Fels made his first report as Secretary and Treasurer to the AEA

Executive Committee on December 27, 1970. And whereas today, he makes his last report as Treasurer of the Association. And whereas seventeen Executive Committees and Presidents have benefited from his guidance, advice and good judgement. For 17 years he has carried out his duties ably and responsibly. We have been privileged to have had his company and counsel for so long. He has served the Association with great distinction and can depart in the sure knowledge of a job well done. He has dealt with a disparate of group of economists with courtesy, patience, and good humor.

Therefore, be it resolved, that the Executive Committee, on behalf of the Association, express its gratitude and appreciation for Ren with a standing round of applause.

The Committee responded enthusiastically.

*Report of the Treasurer* (Fels). Fels reviewed the status of the Ad Hoc Committee on Investments report. The committee has recommended that the Association move its portfolio to Wells Fargo and invest in passively managed, indexed funds. The Finance Committee would review the report and Stein Roe's response and report to the March meeting. No action was needed at the moment.

Fels asked for action on four matters contained in his report (see elsewhere in this issue). It was VOTED to approve an increase in AEA dues to COSSA from \$35,000 to \$38,000. It was VOTED to increase the basic dues rate from \$38.50 to \$42.00 effective January 1, 1989. It was VOTED to approve the 1988 budget (see the Treasurer's Report). And it was VOTED to appoint a special committee to review and evaluate the *Journal of Economic Perspectives* and to recommend a permanent financing arrangement for it. It was understood that the committee would specifically consider the "unbundling" issue—allowing members to select among the three journals with an appropriate price charged for each. Total dues would depend on the number of journals chosen. The committee is expected to report to the December meeting.

*Report of the Committee on the Status of Women in the Economics Profession* (Gordon). Gordon, chair of CSWEP, reported that the committee had published *Women in Economics*, a roster of women economists containing information such as employer, educational background, fields of specialization, and number of publications.

The project to examine differences in career paths of men and women with Ph.D.s in economics made little progress in 1987

because of difficulties regarding access to confidential data maintained by the National Academy of Sciences. These difficulties have been resolved and results should be available in 1988. The Committee's written report is printed elsewhere in this issue.

The meeting adjourned at 4:15 P.M.

Respectfully submitted,  
C. ELTON HINSHAW, *Secretary*

## Report of the Secretary for 1987

*Annual Meeting.* The annual meeting will be held in New York in 1988 and in Atlanta in 1989. Each of these meetings is scheduled for December 28-30 and each will have a Placement Service, which will open for business one day earlier (December 27) than the meetings.

The results of the preferential ballot concerning meeting dates is given in the following table:

	Votes	Rank
Strongly prefer December 27-30	1,022	2
Strongly prefer the 1st weekend after January 1	2,103	1
Prefer December 27-30	611	5
Prefer the 1st weekend after January 1	896	3
Indifferent	758	4

*Elections.* In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee and the Electoral College.

The Nominating Committee, consisting of Charles L. Schultze, Chair, Marjorie Honig, Stephen P. Magee, Frederick Pryor, Michael Rothschild, Richard L. Schmalensee, and Frank P. Stafford submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

### *President-Elect*

Joseph A. Pechman

### *Vice-President*

Martin Feldstein

Roy Radner

F. M. Scherer

A. Michael Spence

### *Executive Committee*

George A. Akerlof

William A. Brock

Edward M. Gramlich

Isabel V. Sawhill

The Secretary prepared biographical sketches of the candidates and distributed ballots last summer. On the basis of the canvass of ballots, I certify that the follow-

ing persons have been duly elected to the respective offices:

President-elect (for a term of one year)

Joseph A. Pechman

Vice-Presidents (for a term of one year)

Martin Feldstein

F. M. Scherer

Executive Committee (for a term of three years)

George A. Akerlof

Isabel V. Sawhill

In addition, I have the following information:

Number of legal ballots	5,804
Number of invalid envelopes	300
Number of envelopes received after October 1	70
Number of envelopes returned	6,174

In accordance with the action taken by the Executive Committee at its March 20, 1987 meeting, amendments to Article III, Sections 2 and 3, and Article IV, Section 7, of the bylaws was submitted to the members in a mail ballot in conjunction with the balloting for officers. I certify that the amendment was approved by a vote of 4,689 "for," 251 "against." The bylaws as amended now read:

### Article III

Section 2. The Association shall have the following officers who shall be appointed by the Executive Committee: a Secretary, a Treasurer, the Editors of its scholarly journals, and a Counsel. The terms of office of each of these officers shall be three calendar years.

Section 3. The Executive Committee shall consist of the President, the President-elect, two Vice-Presidents, the Secretary, the Treasurer, the Editors, the two ex-Presidents who last held office, and six elected members, provided the Secretary, Treasurer, and the Editors shall not be entitled to vote in the Executive Committee's meetings.

### Article IV

Section 7. The Editors shall, with the advice and consent of the Executive Committee, appoint members of Edi-

torial Boards to assist them. The Editors shall be *ex officio* members and chairpersons of their respective Boards, which shall have charge of the publications.

**Membership.** The total number of members and subscribers is shown in Table 1. The total has fluctuated between 25,000 and 26,500 since 1975 when it reached an all-time high of 26,787.

TABLE 1—MEMBERS AND SUBSCRIBERS  
(END OF YEAR)

	1985	1986	1987
Class of Membership			
Annual	17,602	17,148	17,230
Junior	1,670	1,632	1,549
Life	359	358	357
Honorary	30	33	31
Family	472	457	448
Complimentary	473	478	477
Total Members	20,606	20,106	20,092
Subscribers	5,852	5,846	5,748
Total Members and Subscribers	26,458	25,952	25,840

**National Registry.** The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service. Economists looking for jobs and employers are urged to register. This is a placement service that maintains the anonymity of employers. The Association is indebted to the Registry for assistance and supervision at the employment service provided at the annual meetings. Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists*, and their professional obligation to list their openings.

**Permission to Reprint and Translate.** Official permission to quote from, reprint, or translate and reprint articles from the *American Economic Review*, *Journal of Economic Literature*, and the *Journal of Economic Perspectives* totaled 393 in 1987, compared to 280 in 1986. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to obtain the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Associ-

ation suggests that authors charge a fee of \$150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

**AEA Staff.** Mary Winer, Kimberly Adair, Norma Ayres, Dana Coleman, Violet Sikes, and Jacquelyn Woods handle the day-to-day operations of the Association; Marlene Hall organizes the operation of the annual meeting. Their dedication and efficiency make the job of the Secretary tolerable. I wish to express my great gratitude for the excellent work they continue to do.

**Committees and Representatives.** Listed below are those who served the Association during 1987 as members of committees or representatives. The year in parentheses indicates the final year of the term to which they were appointed. On behalf of the Association, I thank them all for their services.

**Ad Hoc Committee on Investments**

Steve Ross, Chair  
Burton M. Malkiel  
Hans R. Stoll

**Budget Committee**

Rendigs Fels, Chair  
Daniel McFadden (1987)  
Sherwin Rosen (1988)  
Judith Thornton (1989)  
Gary S. Becker, *ex officio*

**Census Advisory Committee**

Victor Zarnowitz, Chair (1987)  
Rosanne E. Cole (1987)  
Ben E. Laden (1987)  
Timothy F. Bresnahan (1988)  
Robert J. Genetski (1988)  
Joel Popkin (1988)  
Margaret C. Simms (1988)  
Ben S. Bernanke (1989)  
Dennis W. Carlton (1989)

**Committee on Economic Education**

W. Lee Hansen, Chair (1987)  
Michael K. Salemi (1987)  
William B. Walstad (1987)  
Bruce R. Dalgard (1988)  
Kalman Goldberg (1988)  
Phillip Saunders, Jr. (1988)  
Rendigs Fels, *ex officio*



*Committee on Federal Economic Statistics*

F. Thomas Juster, Chair (1988)  
 Barry P. Bosworth (1988)  
 John Cogan (1988)  
 Rosanne Cole (1988)  
 Ivan Fellighi (1988)  
 Lyle E. Gramley (1988)  
 Zvi Griliches (1988)  
 William A. Morrill (1988)  
 Eugene Smolensky (1988)  
 Robert Solomon (1988)  
 John Wilson (1988)  
 Nina W. Cornell, *ex officio* (COSSA rep.)  
 Marilyn Moon, *ex officio* (COPAFS rep.)  
 Joel Popkin, *ex officio* (COPAFS rep.)  
 Alice M. Rivlin, *ex officio*  
 Gary S. Becker, *ex officio*

*Finance Committee*

Rendigs Fels, Chair  
 Robert G. Dederick (1987)  
 Robert Eisner (1988)  
 Robert J. Genetski (1989)  
 C. Elton Hinshaw, *ex officio*

*Committee on Honorary Members*

Richard E. Caves, Chair (1988)  
 Franco Modigliani (1988)  
 J. Carter Murphy (1990)  
 Gordon C. Winston (1990)  
 Arnold C. Harberger (1992)  
 Anne O. Krueger (1992)

*Committee on Honors and Awards*

Oliver E. Williamson, Chair (1991)  
 John B. Taylor (1987)  
 William Vickrey (1987)  
 James J. Heckman (1989)  
 Richard R. Nelson (1989)  
 Dale W. Jorgenson (1991)

*1987 Nominating Committee*

Charles L. Schultze, Chair  
 Marjorie Honig  
 Stephen P. Magee  
 Frederic Pryor  
 Michael Rothschild  
 Richard L. Schmalensee  
 Frank P. Stafford

*Committee on Political Discrimination*

Burton A. Weisbrod, Chair (1989)  
 Benjamin J. Cohen (1987)  
 Clark W. Reynolds (1987)  
 Lester C. Thurow (1988)  
 Steve Marglin (1989)  
 Stephen E. Margolis (1989)

*Committee on the Status of Minority Groups in the Economics Profession*

Ronald L. Oaxaca, Chair (1988)  
 Bernard E. Anderson (1987)  
 William A. Darity (1987)  
 Glenn C. Loury (1987)  
 Rhonda Williams (1987)  
 George J. Borjas (1988)  
 Vernon J. Dixon (1988)  
 Clifford E. Reid (1988)  
 Margaret C. Simms (1988)

*Committee on the Status of Women in the Economics Profession*

Isabel V. Sawhill, Chair (1987)  
 Karen Davis (1987)  
 Helen Junz (1987)  
 Beth E. Allen (1988)  
 Alan E. Fechter (1988)  
 Nancy M. Gordon (1988)  
 Katharine C. Lyall (1988)  
 Cecilia Conrad (1989)  
 Audrey Freedman (1989)  
 Shulamit Kahn (1989)  
 Judith R. Lave (1989)  
 Marjorie B. McElroy (1989)  
 Gary S. Becker, *ex officio*  
 Joan G. Haworth, Membership Secretary

*AEA/SSRC Joint Committee on U.S.-China Exchanges*

Gregory C. Chow, Co-Chair  
 Kenneth Arrow  
 Lawrence R. Klein  
 Theodore W. Schultz

*Committee on U.S.-Soviet Exchange*

Franklyn D. Holzman, Chair (1987)  
 Jennifer R. Reinganum (1986)  
 Lloyd G. Reynolds (1986)  
 Abram Bergson (1987)  
 Joseph A. Pechman (1988)  
 Richard N. Rosett (1988)

## COUNCIL AND OTHER REPRESENTATIVES

*American Association for the Advancement of Science, Section K, Social Economics and Political Sciences*

Adam Rose (1988)

*American Association for the Advancement of Slavic Studies*

Joseph Brada (1989)

*American Council of Learned Societies*

C. Elton Hinshaw (1990)

*Review Board of the American Statistical Association-Bureau of Census*

Zvi Griliches

*Review Board of the American Statistical Association (ASA)/Bureau of Labor Statistics (BLS)—Research Fellowship and Associate Program*

Robert A. Pollak

*Consortium of Social Science Associations (COSSA)*

Nina W. Cornell (1988)

C. Elton Hinshaw

*Council of Professional Associations on Federal Statistics (COPAFS)*

Marilyn Moon (1988)

Joel Poplin (1988)

*International Economic Association*

Kenneth Arrow (1990)

C. Elton Hinshaw

*Policy Board of the Journal of Consumer Research*

Louis L. Wilde (1988)

*National Bureau of Economic Research*

David A. Kendrick (1987)

*National Council for Social Studies*

W. Lee Hansen

*Social Science Research Council*

Hugh T. Patrick (1987)

## REPRESENTATIVES OF THE ASSOCIATION ON VARIOUS OCCASIONS—1987

*Inaugurations*

Albert Emanuel Smith, South Carolina State College

William H. Wesson, Jr.

M. Christopher White, Gardner-Webb College

Stanley J. Dudko

C. ELTON HINSHAW, *Secretary*

# Report of the Treasurer for the Year Ending December 31, 1987

The budget for 1988 approved by the Executive Committee on December 27, 1987, is shown in Table 1. It projects an operating loss of \$514 thousand, an investment gain of \$263 thousand, and an overall deficit of \$251 thousand. The projected ratio of net worth at the beginning of 1988 to budgeted expenses for that year is 1.44. Since a ratio of 1.0 is deemed adequate, the finances of the Association are in good shape in spite of the deficit. The gyrations of the stock market in

1987 have had little net effect on the portfolio of the Association.

The deficits for 1987 and 1988 resulted from the decision to begin publication of the new *Journal of Economic Perspectives*. The Executive Committee originally appropriated half a million dollars for its start-up costs. The Committee decided to wait until the journal proved successful before deciding on permanent financing for it. Although the original appropriation will be exceeded dur-

TABLE 1—1988 BUDGET, AMERICAN ECONOMIC ASSOCIATION: Approved December 27, 1987  
(thousands of dollars)

	First Nine Months (Unaudited)		Actual	Full Year Budgeted	
	1986	1987		1986	1987
<b>REVENUES FROM DUES AND ACTIVITIES</b>					
Membership dues	\$647	675	870	\$870	900
Nonmember subscriptions	483	488	650	680	664
Subtotal	1,130	1,163	1,520	1,550	1,564
Subscriptions, <i>Job Openings for Economists</i>	21	22	32	30	33
Advertising	95	102	130	130	140
Sale of <i>Index of Economic Articles</i>	9	61	43	150	200
Sales of copies, republications, handbooks	30	28	40	27	27
Sale of mailing list	30	33	51	46	55
Annual meeting	37	59	37	16	50
Sundry	48	59	73	64	80
<b>Total Operating Revenue</b>	<b>1,400</b>	<b>1,527</b>	<b>1,926</b>	<b>2,013</b>	<b>2,149</b>
<b>PUBLICATION EXPENSES</b>					
<i>American Economic Review</i>	457	469	613	642	686
<i>Journal of Economic Literature</i>	582	609	768	852	911
<i>Journal of Economic Perspectives</i>	36	152	75	278	358
Directory	53	53	70	70	70
<i>Job Openings for Economists</i>	36	38	55	57	60
<i>Index of Economic Articles</i>	6	41	31	75	100
Subtotal	1,170	1,361	1,611	1,974	2,185
<b>OPERATING AND ADMINISTRATIVE EXPENSES</b>					
General and Administrative	220	222	337	350	368
Committees	27	23	45	50	55
Support of other organizations	49	40	62	52	55
Subtotal	296	284	444	452	478
<b>Total Expenses</b>	<b>1,466</b>	<b>1,645</b>	<b>2,055</b>	<b>2,426</b>	<b>2,663</b>
OPERATING GAIN (LOSS)	(66)	(118)	(129)	(413)	(514)
INVESTMENT GAIN (LOSS)	336	203	248	271	263
SURPLUS (DEFICIT)	270	85	119	(142)	(251)
<b>Ratio, Net Worth to Annual Expenses</b>			<b>1.70</b>	<b>1.64</b>	<b>1.44</b>

ing 1988, there is no need for immediate action other than the increase in dues scheduled for January 1, 1989. The base rate will go up from \$38.50 to \$42.00 with other dues rates going up in proportion. This will not prevent another deficit in 1989, but it is expected to keep the ratio of net worth to expenses above the 1.0 target ratio. Further action to increase revenues will be needed for 1990.

Audited financial statements for 1987 will be published in the June issue of the *American Economic Review*.

In submitting my last report after seventeen years as Treasurer, I want to express my gratitude to the staff in the Nashville office, present and past, especially Norma Ayres, the best accountant a treasurer could hope for; Mary Winer, who took over as Administrative Director when I was still Secretary-Treasurer and has operated the office with extraordinary efficiency; and Violet Sikes, who while carrying out other duties has helped me in many ways.

RENDIGS FELS, *Treasurer*

## Report of the Finance Committee

The Finance Committee of the American Economic Association met at the Chicago Club, Chicago, Illinois, at noon on December 11, 1987. Present were Robert Dederick, Robert Eisner, Rendigs Fels (Chairman), and Robert Genetski (members of the Committee); Harvey Hirschhorn, Robert McNeill and James Weiss (representing Stein Roe & Farnham, Investment Counsel for the Association); and C. Elton Hinshaw (Treasurer-elect).

Copies of the Report of the AEA Committee on Indexing Association Funds dated December 9, 1987, were distributed. This Report recommends switching the Association's funds from active management by Stein Roe & Farnham to the family of passive index funds managed by Wells Fargo Investment Advisors. Discussion of the Report identified a number of issues:

(1) What are the portfolio objectives of the Association? Specifically, is the object to match (or beat) some index such as the S & P 500, or is it to protect the real value of the portfolio?

(2) Should the funds be managed by Stein Roe, Wells Fargo, or someone else?

(3) What are the costs of the alternatives?

(4) Should management of the funds be active or passive?

(5) How should decisions about allocating funds (for example, between equities and bonds) be made?

On (1), the Finance Committee felt that the Executive Committee should address the question of the objectives the Association seeks in portfolio management. Specifically the Investment Counsel has been operating since the 1960's on instructions that maintaining the real value of the portfolio is the main consideration. At one time, the policy of the Association was to put everything into equities to protect against inflation. Until now, maximixing the rate of return has not been the primary consideration. If the Executive Committee wants the Investment Counsel to match or beat the return on the S & P 500, it should be explicit about the objective.

On (2), the Finance Committee had too little knowledge of Wells Fargo to pass judgment on it but felt that Stein Roe & Farnham had done a good job.

On (3), the Finance Committee found the calculations on page 8 of the Report of the AEA Committee on Indexing Association fund erroneous. (The error was subsequently corrected by the Committee on Indexing.)

On (4), two members of the Finance Committee thought that active management of funds was preferable to meet changing conditions. Eisner favored passive management.

On (5), it was not clear how decisions allocating funds between equities, bonds and cash would be made under passive management. The current directive to Stein Roe calling for 50 to 75 percent in equities is appropriate for a professional Investment Counsel but too broad if allocation were to be entrusted to the Treasurer. A Finance Committee that meets once a year, as has been the case since 1970, could not specify a narrow range. This problem needs to be studied carefully.

Since the representatives of Stein Roe wanted to submit a counterproposal to the Wells Fargo option, the following procedure was decided by Hinshaw and myself. The counterproposal will be mailed to the members of the Finance Committee along with the revised report of the Committee on Indexing. Their views together with the report and the counterproposal will be submitted to the Executive Committee for decision at its March 1988 meeting.

In addition to discussing indexing, the Finance Committee took up the question of the 50-75 equities directive. Stein Roe had reduced the actual proportion to barely more than 50 percent before the stock market crash, which carried the proportion well below that figure. The Committee decided not to require Stein Roe to buy equities to restore the actual proportion to the 50-75 percent range.

Members can obtain a list of the assets in the portfolio by writing the Treasurer.

RENDIGS FELS, *Chairman*

# Report of the Editor

## American Economic Review

The editorial process has worked very smoothly during the past year despite several changes in personnel at the *Review* that I outline below. As I indicated in my report of last year, the major effect of the new editorial process has been to increase the editorial effort devoted to submitted manuscripts, and this has caused some slowdown in the speed with which manuscripts are processed.

### Editorial Process

In order to try to speed up the processing of manuscripts we have implemented two changes in our editorial procedures. First, some of the papers submitted to the *Review* are clearly not in a form ready for editorial judgment. Every managing editor of the *Review* in the last twenty years has complained about this problem. To try to remedy it, I proposed raising the submission fee (to \$50 for AEA members, the vast majority of those submitting papers, and to \$100 for non-members), and this was approved by the AEA Executive Committee and went into effect on May 1, 1987. In fact, the real value of the *Review's* submission fee had declined considerably in the last few years, so the increase was long past due. It is my hope and the hope of my co-editors that this change will reduce the number of manuscripts that are inappropriately submitted for consideration by the *Review*.

As Table 1 indicates, the number of submissions has declined from last year. Since we are publishing fewer papers than we have in the past, this has meant that the chances that a submitted paper will eventually be published have remained constant from last year to this year.

Although it is tempting for an economist to do so, there is no strong evidence to support the view that an increased submission fee is responsible for the decline in submissions to the *Review*. As it turns out,

TABLE 1—MANUSCRIPTS SUBMITTED  
AND PUBLISHED, 1968–87<sup>a</sup>

Year	Submitted	Published	Ratio Published-to-Submitted
1968	637	93	.15
1969	758	121	.16
1970	879	120	.14
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15
1974	723	125	.17
1975	742	112	.15
1976	695	117	.17
1977	690	114	.17
1978	649	108	.17
1979	719	119	.17
1980	641	127	.20
1981	784	115	.15
1982	820	120	.15
1983	932	129	.14
1984	921	138	.15
1985	952	128	.13
1986	987	123	.125
1987	843	99	.12

<sup>a</sup>The submissions reported for every year refer to the last two months of the previous year and the first ten months of the year reported.

we had already observed a decline in the number of submissions to the *Review* in the beginning of 1987, before the increase in the submission fee, so there are certainly other forces at work.

Table 2 indicates that we published fewer papers and pages of material in the *Review* in 1987 than in 1986. This is consistent with the policy we have had of increasing the size of the *Review* over the period since 1984 to deliberately reduce the large backlog of accepted papers that I inherited from my predecessor. The number of pages published in the *Review* was greater in 1985 and 1986 than in 1984, while the 1987 figures reflect a return to the experience of 1984 now that our backlog has been reduced. Although the number of articles has now stabilized at about its 1984 level, our publication of notes, comments, and replies has decreased steadily

TABLE 2—SUMMARY OF CONTENTS, 1986 AND 1987

	1986		1987	
	Number	Pages	Number	Pages
Articles	55	803	50	746
Shorter Papers, including Comments and Replies	66	344	47	272
Dissertations		21		—
Announcements and Notes Section		53		56
Index		10		9
Total		1268		1083

since 1985. Both I and my co-editors believe this is a desirable editorial change. Our goal is to increase the number of major, important research papers published in the *Review*, and we expect this to come mainly at the expense of our publication of very brief notes and comments.

Tables 3 and 4, when compared with the results for last year, indicate there has been no change in the speed with which we handle manuscripts that are ultimately rejected. In my view, and the view of my co-editors, these data indicate too many long delays in the editorial decision-making process.

A major cause of the delay in handling manuscripts can be attributed, as I indicated in my Report last year, to the increased use of referees and to the slowness with which we receive reports. One way to speed up our editorial process would therefore be to speed up our receipt of referee reports.

TABLE 3—DISPOSITION OF MANUSCRIPTS, 1986 AND 1987

	July 1, 1985– June 30, 1986	July 1, 1986– June 30, 1987
Manuscripts Received	982	897
Completed Processing	468	609
Accepted	31	23
Rejected	437	586
Currently in Process	514	288

In line with this view, I proposed to the Executive Committee that we pay referees a fee of \$35 if they produced a report for us within four weeks. The Committee approved my request and we have been paying referees for promptly received reports since May 1, 1987. Coupled with our increased submission fee, I also indicated that this change

TABLE 4—DISTRIBUTION OF EDITORIAL DECISION LAGS BETWEEN RECEIPT AND REJECTION, JULY 1, 1986–JUNE 30, 1987

Weeks to Rejection	Total Number of Manuscripts	Percent	No Outside Referees	1 Referees	2 Referees	3 or More Referees
0–4	55	.07	38	13	3	1
5–6	36	.05	8	15	11	2
7–8	64	.09	8	29	20	7
9–10	64	.09	8	27	16	13
11–12	68	.09	5	22	30	11
13–14	62	.08	3	19	26	14
15–16	47	.06	2	17	19	9
17–21	113	.15	4	38	54	17
22–26	80	.11	1	27	40	12
27–30	60	.08	2	20	23	15
31–35	31	.04	0	9	15	7
36–52	63	.09	2	2	53	6
	743	100.00	81	238	310	114

TABLE 5—AVERAGE PUBLICATION LAGS,  
BY JOURNAL ISSUE

Issue	Number of Weeks Lag		
	Receipt to Acceptance	Acceptance to Publication	Receipt to Publication
March 1987	51	21	72
June 1987	49	16	65
September 1987	59	18	77
December 1987	67	23	90

would probably be revenue neutral, as it appears to have been.

The purpose of making payments to referees is simple: It is designed to increase the speed with which referees respond to requests at the *Review*. It is far too early to tell whether this change will increase the speed with which we can process manuscripts, but the results to date are very encouraging.

Table 5 indicates that there has been no change from last year in the speed with which accepted papers are processed and published in the *Review*. (The December 1987 figure is an aberration that we do not expect will be repeated.)

The subject matter distribution of papers published in the *Review* in 1986 and 1987 is contained in Table 6. It remains my impression that the distribution of published papers

TABLE 6—SUBJECT MATTER DISTRIBUTION OF  
PUBLISHED MANUSCRIPTS, 1986 AND 1987

	Published	
	1986	1987
General Economics and General Equilibrium Theory	9	4
Microeconomic Theory	13	12
Macroeconomic Theory	12	8
Welfare Theory and Social Choice	1	3
Economic History, History of Thought, Methodology	7	1
Economic Systems	1	1
Economic Growth, Development, Planning, Fluctuations	6	2
Economic Statistics and Quantitative Methods	2	2
Monetary and Financial Theory and Institutions	6	6
Fiscal Policy and Public Finance	18	9
International Economics	8	8
Administration, Business Finance	4	0
Industrial Organization	14	25
Agriculture, Natural Resources	1	4
Manpower, Labor Population	13	7
Welfare Programs, Consumer Economics, Urban and Regional Economics	6	7
Total	121	99

reflects fairly accurately the distribution of papers submitted.

### Papers and Proceedings

The tenth volume of the *Papers and Proceedings* to be prepared by the editorial staff of the *Review* appeared in May 1987. This task was handled by Harvey Rosen (of

TABLE 7—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING, 1987 *AER*

	Copies Printed	Pages		Cost		
		Net	Gross	Issue	Reprints	Total
March	27,500	242	288	\$50,099.59	\$1,106.10	\$51,206
May	28,000	404	432	68,033.82	2,443.30	70,477
June	27,500	266	296	52,561.82	1,294.52	53,856
September	27,500	278	328	57,225.53	1,183.26	58,409
December <sup>a</sup>	27,500	284	346	59,166.00	2,100.00	61,266
Annual Misc. <sup>b</sup>						10,000
Total		1,474	1,690	\$287,086.76	\$8,127.18	\$305,214

<sup>a</sup>Estimated.

<sup>b</sup>Estimated: based on costs of preparing mailing list, extra shipping charges, and storage costs of back issues.



Princeton University) and Wilma St. John. I am deeply indebted to both of them for the difficult work under extraordinarily tight deadlines that they have so capably performed.

### Co-Editors and Board of Editors

There have been several major changes in the editorial personnel at the *Review* in the last year. John Riley resigned his position as co-editor at the *Review* in the summer of 1987. I am deeply grateful for all the help and effort he provided over his two-year period as co-editor. Some aspects of editing the *Review* necessarily involve unpleasantness since, after all, a major part of the post involves delivering unpleasant news to potential authors. For me, a very pleasant aspect of editing the *Review* is the close working relationship that has evolved among the co-editors. It has been a special pleasure to know John Riley in this way.

Hal Varian (of the University of Michigan) graciously consented to take Riley's place as co-editor and so I now edit the *Review* with his assistance as well as that of Robert Haveman and John Taylor. I am deeply indebted to them for the conscientious effort they have expended over the last year.

The Board of Editors now consists of sixteen members and I am indebted to them also for their efforts. Board members are selected to reflect the highest level of scholarship in the economics profession from the breadth of different fields represented in our submissions. More than fine scholarship is expected of a Board member, however. Board members are also selected because of their conscientiousness, good judgment, and professional reliability. When possible, we like to select Board members from those econo-

mists who have been especially helpful in the outside refereeing process.

Seven members of the Board completed their terms as of March 31, 1987: Clive Bull, Michael Darby, Philip Graves, Meir Kohn, Paul Krugman, Susan Woodward, and Leslie Young. I am most grateful to them and to the continuing members: George Akerlof, Jacob Frenkel, Claudia Goldin, Jo Anna Gray, George Johnson, John Kennan, Mervyn King, Bennett McCallum, Maurice Obstfeld, Edgar Olsen, Robert Porter, Richard Roll, Alvin Roth, Steven Shavell, John Shoven, and Kenneth Singleton.

Wilma St. John retired from her position of production editor this year after nearly nineteen years in that position. I and two previous managing editors are deeply grateful for her many efforts over the years. I am personally grateful for her willingness to set up our new offices in Princeton and for her dedication to the highest standards for the *Review*. Although Wilma has left the post of production editor, she will continue to edit the *Papers and Proceedings* with Harvey Rosen. Claire H. Comiskey has taken on the role of production editor and I am indebted to her and to our office manager Shirley Griesbaum and our editorial assistant Sandra Grant for the fine work they have continued to perform over the past year. I would also like to thank the co-editors' secretaries: Marsha Stanik (Robert Haveman's office); Alice Ann Halvorsen and Jamie Sidells (John Taylor's office); Beverly Bardi and LeTeen Fogelquist (John Riley's office); and Carolyn Bartle (Hal Varian's office).

The published version of this report contains the list of referees who have volunteered their services in 1987. We extend our deepest appreciation for the time and energy they have devoted to the advancement of our science.

A. B. Abel  
J. M. Abowd  
K. G. Abraham  
P. Aghion  
G. Akerlof  
J. W. Albrecht  
A. Alchian

H. Alderman  
A. Alesina  
B. Allen  
F. Allen  
S. G. Allen  
M. Allingham  
J. Alm

J. G. Altonji  
E. Alvi  
C. An  
J. Anderson  
J. E. Anderson  
A. Ando  
J. Andreoni

J. J. Antel  
R. N. Anthony  
E. Appelbaum  
R. J. Arnott  
K. Arrow  
W. B. Arthur  
C. Ash

- |                 |                   |                  |                 |
|-----------------|-------------------|------------------|-----------------|
| R. Ashley       | A. S. Blinder     | C. D. Bull       | P. Coughlin     |
| S. Atkinson     | C. Bliss          | R. Burkhauser    | P. N. Courant   |
| A. J. Auerbach  | C. Blitzler       | G. T. Burtless   | D. L. Coursey   |
| C. Azzi         | M. K. Block       | J. S. Butler     | F. Cowell       |
| D. K. Backus    | D. E. Bloom       | R. J. Butler     | T. Cowen        |
| K. Bagwell      | L. Blume          | C. Calomiris     | T. G. Cowing    |
| M. J. Bailey    | R. Blundell       | R. Calvert       | J. C. Cox       |
| R. Baldwin      | R. Boadway        | G. Calvo         | A. Craig        |
| L. Ball         | Z. Bodie          | E. Cameron       | S. G. Craig     |
| R. D. Banker    | D. Bohi           | T. Cameron       | P. Cramton      |
| J. Banks        | L. Boland         | C. Campbell      | R. W. Crandall  |
| P. K. Bardhan   | P. Bolton         | J. Y. Campbell   | J. Craven       |
| W. A. Barnett   | R. Bolton         | J. Campos        | V. Crawford     |
| D. Baron        | E. Bomhoff        | A. Caplan        | K. Crocker      |
| R. J. Barro     | C. Bond           | A. Caplin        | M. Cropper      |
| J. B. Barron    | J. Bonin          | D. Card          | J. Cuddington   |
| R. Barsky       | M. Bordo          | D. W. Carlton    | A. Cukierman    |
| A. Bartel       | S. Borenstein     | L. H. Carmichael | J. Culbertson   |
| G. Basevi       | G. J. Borjas      | J. Caskey        | R. Cumby        |
| L. J. Bassi     | D. Bos            | R. Caves         | R. Cummings     |
| T. Bates        | J. Boschen        | G. Chamberlin    | C. Dahlman      |
| C. Baum         | M. Boskin         | E. C. Chang      | R. E. Dansby    |
| W. J. Baumol    | B. Bosworth       | S. Chaplinsky    | S. Danziger     |
| M. Baxter       | B. L. Boulier     | K. Chatterjee    | P. Danzon       |
| C. M. Beach     | L. Bovenberg      | M. Cherkes       | M. R. Darby     |
| B. Beatty       | S. Bowles         | H. Chernick      | J. Darmstadter  |
| L. Bebchuk      | R. Boyer          | A. Chesher       | P. A. David     |
| G. S. Becker    | K. Bradbury       | S. Cheung        | C. Davidson     |
| W. Becker       | D. Bradford       | J. Chipman       | J. Davies       |
| D. Begg         | R. M. Bradford    | B. Chiswick      | J. C. Dawson    |
| J. A. Bell      | M. Bradley        | G. C. Chow       | R. Day          |
| J. Benhabib     | W. Brainard       | C. Christ        | A. V. Deardorff |
| Y. Benjamini    | S. D. Braithwait  | L. Christensen   | A. Deaton       |
| Y. Ben-Porath   | B. Branch         | L. Christiano    | H. de Gorter    |
| G. Bentson      | J. A. Brander     | C. C. Y. Chu     | P. DeGraba      |
| A. Berger       | L. Brandt         | K. Clark         | H. Dellas       |
| E. Berglas      | W. H. Branson     | P. Clark         | B. DeLong       |
| A. Bergson      | G. Brennan        | D. L. Cleeton    | N. Demarchi     |
| T. Bergstrom    | T. F. Bresnahan   | C. Clotfelter    | H. Demsetz      |
| B. S. Bernanke  | A. A. Brewer      | J. Cochrane      | E. Denison      |
| E. R. Berndt    | M. Bronfenbrenner | S. Collins       | M. Denny        |
| D. Bernheim     | D. Brookshire     | W. S. Comanor    | A. Denzau       |
| C. H. Berry     | C. C. Brown       | D. Con           | A. Deolalikar   |
| S. Berry        | G. Brown          | J. Conlisk       | P. Desai        |
| H. Bester       | J. N. Brown       | P. J. Cook       | A. Devany       |
| J. Bhagwati     | E. Browning       | T. F. Cooley     | W. Dewald       |
| S. Bhattacharya | M. J. Browning    | R. W. Cooper     | P. Diamond      |
| N. Birdsall     | N. Bruce          | R. Cooter        | B. Diba         |
| J. Bishop       | J. K. Brueckner   | W. Corden        | W. T. Dickens   |
| R. Bishop       | J. Bryant         | J. J. Cordes     | E. Diewert      |
| O. J. Blanchard | R. Bryant         | M. Cordomi       | A. K. Dixit     |
| M. Blaug        | E. Buffie         | B. Cornell       | D. Dollar       |

D. Donaldson	T. Flaherty	W. J. Gormley	J. A. Hausman
J. Donaldson	M. A. Flavin	P. Gottschalk	T. Havrilesky
R. Dorfman	R. P. Flood	D. Granick	J. Hay
R. Dornbusch	R. Fogel	P. E. Graves	F. Hayashi
M. Dotsey	D. Foley	J. A. Gray	J. Head
W. R. Dougan	J. E. Foster	W. H. Greene	G. M. Heal
T. Downs	R. Frank	M. Greenhut	J. J. Heckman
A. Drazen	H. Fräzis	G. S. Greenslade	A. Heertje
J. Driffill	T. Frech III	P. Gregory	P. Helmberger
R. Driskill	M. Freeman	G. Grenier	P. Hendershott
K. B. Dunn	K. R. French	D. M. Grether	J. V. Henderson
S. Durland	J. Frenkel	J. M. Griffin	Z. Hercowitz
P. Dybvig	A. F. Friedlander	E. Grinols	E. Hewett
J. Eaton	B. Friedman	T. S. Gronberg	A. Hillman
U. Ebert	D. Friedman	G. Grossman	J. R. Hines
L. Ebrill	J. Friedman	M. Grossman	D. Hirschleifer
Z. Eckstein	M. Friedman	P. Grout	J. Hirschleifer
B. Eden	I. Friend	L. Guasch	J. Hoadley
L. Edwards	K. Froot	R. Guesnerie	I. Hoch
S. Edwards	D. D. Fudenberg	J. Gyourko	R. Hodgkin
R. Ehrenberg	M. A. Fuss	P. Haaparanta	R. Hodrick
I. Ehrlich	M. Gaffney	F. Hahn	A. S. Holland
M. S. Eichenbaum	J. Gagnon	R. Hahn	S. Hollander
J. Enelow	D. Gale	W. J. Haley	B. Holmstrom
C. Engel	I. Gale	B. H. Hall	C. A. Holt
S. Engerman	H. Galper	J. C. Haltiwanger	D. Holtz-Eakin
M. Engers	P. Garber	D. Hamermesh	F. Holzman
D. Epple	J. Geanakopulos	B. W. Hamilton	P. Hooper
L. Epstein	J. Genoska	J. Hamilton	B. Horrigan
R. Ericson	V. Geraci	P. Hammond	J. V. Hotz
N. Ericsson	J. D. Germany	D. Hancock	E. P. Howrey
S. Estrin	M. Gersovitz	T. H. Hannan	W. E. Hoyt
W. Ethier	M. Gertler	G. D. Hansen	D. Hsieh
G. W. Evans	P. Gertler	L. Hansen	S. Hu
P. Evans	R. Gertner	R. G. Hansen	F. M. Huang
R. C. Fair	J. Geweke	H. Hansmann	R. G. Hubbard
E. F. Fama	F. D. Gianfrancesco	I. Hansson	J. Huizinga
H. S. Farber	M. R. Gibbons	E. Hanushek	C. R. Hulten
J. Farrell	R. S. Gibbons	A. C. Harberger	M. D. Hurd
S. Fazzari	R. F. Gillingham	J. Harrigan	A. Hurter
D. Feenberg	A. Giovannini	D. Harris	J. Ingersoll
R. C. Feenstra	M. Gisser	J. Harris	R. P. Inman
E. Feige	V. Goldberg	R. G. Harris	M. D. Intriligator
A. Feldman	A. Goldberger	A. J. Harrison	O. Irvine
M. Feldstein	S. Goldfeld	G. Harrison	R. M. Isaac
J. Ferejohn	C. Goldin	J. C. Harsanyi	P. Isard
G. S. Fields	M. Goldstein	O. D. Hart	Y. Ishii
S. Figlewski	F. M. Gollop	P. Hartley	T. Ito
S. Fischer	M. Goodfriend	R. Hartman	R. Jacobson
A. Fisher	D. M. Gordon	J. Hasbrouch	A. B. Jaffe
F. M. Fisher	M. Gordon	M. Hashimoto	G. H. Jakubson
J. Fitzgerald	R. J. Gordon	J. C. Hause	M. C. Jensen

- |               |                 |                    |                  |
|---------------|-----------------|--------------------|------------------|
| G. Johnson    | R. Kormendi     | P. Lindert         | A. Marcet        |
| M. Johnson    | J. Kornai       | C. M. Lindsay      | A. Marcus        |
| M. C. Johnson | L. Kotlikoff    | C. Link            | D. Margaritis    |
| P. R. Johnson | I. Kravis       | P. Linneman        | R. Margo         |
| R. Johnson    | K. Krishna      | B. Lipman          | N. Mark          |
| T. Johnson    | A. B. Krueger   | S. A. Lippman      | J. R. Markusen   |
| W. R. Johnson | P. Krugman      | S. C. Littlechild  | J. Marquez       |
| J. Jondrow    | J. Krutilla     | R. H. Litzenberger | S. Masters       |
| R. Jones      | F. Kuchler      | M. Loeb            | J. Mather        |
| P. L. Joskow  | P. Kuhn         | K. Lofgren         | F. G. Mathewson  |
| K. Judd       | H. Kunreuther   | J. Long            | R. A. Maynard    |
| J. H. Kagel   | F. Kydland      | C. Los             | C. McLure        |
| J. Kahn       | H. Ladd         | J. R. Lott         | W. H. Meckling   |
| L. M. Kahn    | D. Laidler      | K. Lovell          | P. Meguire       |
| D. Kahneman   | R. J. LaLonde   | D. Lucas           | Y. P. Mehra      |
| J. Kalt       | D. Lam          | R. E. Lucas, Jr.   | G. Meier         |
| M. Kamlet     | D. A. Lam       | R. E. B. Lucas     | A. H. Meltzer    |
| E. Kane       | D. Landau       | M. Lundahl         | J. Melvin        |
| S. H. Karlson | K. Lang         | M. Lundberg        | M. Melvin        |
| D. W. Karlton | G. LaRoque      | S. Lundberg        | J. Meyer         |
| E. Karni      | D. Larsen       | L. Lynch           | L. Meyer         |
| E. Katz       | R. W. Latham    | R. M. Lyon         | R. Michael       |
| L. F. Katz    | L. J. Lau       | R. P. McAfee       | R. W. Michener   |
| M. Katz       | C. Lave         | M. E. McBride      | P. Mieszkowski   |
| J. B. Kau     | L. Lave         | J. McCall          | J. Migue         |
| R. Kaufman    | D. Lavoie       | B. T. McCallum     | H. Milde         |
| S. Kealhofer  | R. Lawrence     | D. N. McCloskey    | P. Milgrom       |
| T. Keeler     | E. E. Leamer    | R. McCullough      | M. Miller        |
| W. Keeton     | T. Lee          | J. B. McDonald     | P. Miller        |
| P. Kehoe      | W. Lee          | J. M. McDowell     | R. A. Miller     |
| M. Kemp       | N. Leff         | M. McGuire         | T. W. Mirer      |
| P. Kemper     | K. Leffler      | G. W. McKenzie     | L. J. Mirman     |
| P. Kenen      | B. N. Lehmann   | M. McMillan        | E. J. Mishan     |
| J. Kennan     | H. Leibenstein  | L. Maccini         | O. Mitchell      |
| R. Kenney     | L. Leiderman    | L. J. Maccini      | P. Mitszvowski   |
| N. Khanna     | P. Leigh        | M. Machina         | H. Miyazake      |
| D. Kiefer     | L. S. Leighton  | R. Mackay          | D. Modest        |
| R. Kihlstrom  | A. Leijonhufvud | J. K. MacKie-Mason | F. Modigliani    |
| E. H. Kim     | H. Leland       | T. E. MaCurdy      | R. A. Moffitt    |
| M. King       | H. Leonard      | A. Maddison        | H. Mohring       |
| R. King       | J. S. Leonard   | J. H. Makin        | J. M. Montias    |
| S. King       | S. Lepper       | G. E. Makinen      | J. Moore         |
| N. Kiyotaki   | S. Leroy        | S. Mallikamas      | E. Morey         |
| B. Klein      | M. D. Levi      | D. H. Malmquist    | S. Morley        |
| P. Klemperer  | R. Levich       | K. J. Maloney      | D. T. Mortenson  |
| M. Knetter    | D. Levine       | P. R. Manger       | T. Mroz          |
| M. Knight     | J. Levinsohn    | N. G. Mankiw       | J. Muellbauer    |
| C. R. Knoeber | T. Lewis        | A. Manne           | D. Mueller       |
| B. Kobayashi  | R. Libby        | R. Manning         | D. J. Mullineaux |
| R. Koenker    | M. Lieberman    | M. Manove          | A. Munnell       |
| M. Kohn       | A. Liebowitz    | E. Mansfield       | R. Murnane       |
| G. Kopits     | R. C. Lind      | R. Manuelli        | K. J. Murphy     |

K. G. Murty	D. Papell	M. R. Ransom	H. E. Ryder, Jr.
R. Musgrave	M. Parkin	P. Rappoport	T. Rymes
J. Mutti	D. O. Parsons	R. H. Rasche	K. Ryoo
S. C. Myers	M. Pascoa	D. J. Ravenscroft	P. Ryscavage
R. Myerson	F. Paukert	E. Ray	E. Sadka
I. M. Nadiri	J. Paulus	A. Razin	B. Sahni
L. Nakamura	M. Pauly	P. Reagan	S. W. Salant
I. C. Nam	C. Paxson	C. E. Reid	G. Saloner
M. P. Narayanan	J. Pechman	R. J. Reilly	S. Salop
J. Nason	A. Peck	C. Reimers	L. Samuelson
J. P. Neary	M. Peck	P. C. Reiss	P. A. Samuelson
J. Neelin	S. Pejovich	J. Reitzes	W. F. Samuelson
J. Neihaus	S. Peltzman	M. Reynolds	T. Sandler
C. R. Nelson	J. Pencavel	D. Richardson	A. Sandmo
J. Nelson	J. F. Perez-Lopez	W. C. Riddell	G. Santoni
M. Nelson	M. Perry	J. Ritter	D. Sappington
R. D. Nelson	I. Persson-Tanimura	J. Ritzen	T. Sargent
E. Nennenhorn	P. Pestieau	F. L. Rivera-Batiz	P. Saunders
R. Netzer	G. Peters	J. Roback	W. Saupe
D. Newbery	B. C. Petersen	D. J. Roberts	W. Scanlon
W. Newey	D. R. Peterson	K. W. S. Roberts	L. C. Schall
Y. Ng	G. Peterson	R. Roberts	T. Schelling
S. Nickell	C. Phelps	P. K. Robins	F. M. Scherer
M. Noether	E. S. Phelps	S. Robinson	G. Schinasi
R. Noll	L. Philips	J. Roemer	A. Schlaifer
W. Nordhaus	R. S. Pindyck	C. Rogers	R. L. Schmalensee
J. R. Norsworthy	C. Pitchik	K. S. Rogoff	D. Schmeidler
D. C. North	M. M. Pitt	V. V. Roley	P. J. Schmidt
N. Noto	M. Plant	R. Roll	R. Schmidt
W. Novshek	G. Plesko	C. D. Romer	P. Schoemaker
J. Nugent	C. Plosser	P. M. Romer	F. G. Schoumaker
J. A. Nyman	R. Plotnick	T. Romer	J. Schroder
W. Oakland	C. R. Plott	A. Rose	M. Schroder
W. E. Oates	I. P'ng	S. Rose-Ackerman	R. Schwab
R. L. Oaxaca	M. Polinsky	S. Rosefields	A. J. Schwartz
M. Obstfeld	R. Pollak	S. Rosen	E. Schwartz
S. Oh	R. Porter	N. Rosenberg	S. Schwartz
K. Ohno	R. C. Porter	M. R. Rosenzweig	G. W. Schwert
H. Ohta	R. H. Porter	J. Rotemberg	S. Scotchmer
W. Y. Oi	P. Portney	A. E. Roth	J. Seade
K. Okuguchi	J. M. Poterba	R. S. Ruback	U. Segal
E. Olson	E. Prescott	P. H. Rubin	P. Segerstrom
S. Oster	A. Protopapadakis	D. Rubinfeld	A. Sen
J. Oströy	F. Pryor	J. Rudin	L. W. Senbet
A. J. Oswald	G. Psacharopoulos	R. J. Ruffin	R. Settle
D. Orr	G. Pyatt	A. M. Rufolo	R. Shakotko
S. Ozler	R. E. Quandt	C. Ruhm	C. Shapiro
H. Pack	J. F. Quinn	M. Rush	M. Shapiro
M. Pagano	R. Radner	C. Russell	P. Shapiro
M. Paglin	D. Rae	T. Russell	S. Sharp
A. Pakes	R. Ram	W. Russell	S. Shavell
J. Panzar	K. Ramaswamy	V. W. Ruttan	S. Sheffrin

- |                  |                   |                     |                  |
|------------------|-------------------|---------------------|------------------|
| D. Shepard       | M. Staszheim      | R. D. Tollison      | W. C. Wheaton    |
| W. Shephard      | R. Steinberg      | S. Tomlinson        | L. H. White      |
| R. Sherman       | J. Steinbrunner   | L. S. Topel         | M. White         |
| R. J. Shiller    | J. Steinhart      | R. Topel            | C. Whiteman      |
| A. Shleifer      | J. Stewart        | R. M. Townsend      | E. R. Wicker     |
| A. Shorrocks     | M. Stewart        | R. Tresch           | J. A. Wilcox     |
| J. Shoven        | G. Stigler        | R. Triest           | D. E. Wildasin   |
| W. F. Shughart   | J. Stock          | R. Tryon            | L. Wilde         |
| D. R. Siegel     | A. Stockman       | J. Tschirhart       | P. Wiles         |
| J. Siegfried     | D. J. Stockton    | G. Tullock          | T. D. Willet     |
| E. Silberberg    | T. M. Stoker      | P. Turner           | A. W. Williams   |
| L. Simon         | M. Strasheim      | D. T. Ulph          | J. Williamson    |
| C. Sims          | J. Strauss        | D. Usher            | O. E. Williamson |
| N. Singh         | R. Strauss        | S. Van Wijnbergen   | S. Williamson    |
| K. Singleton     | C. Stuart         | S. Vassilikas       | R. D. Willig     |
| H. Sinn          | R. Stulz          | J. Veitch           | R. J. Willis     |
| A. Siow          | W. Suen           | Y. Venieris         | C. Wilson        |
| L. Sjaastad      | D. Sullivan       | W. Vickrey          | J. Wilson        |
| J. Slemrod       | A. Summers        | W. P. M. Vijverberg | J. D. Wilson     |
| F. Sloan         | L. Summers        | V. Vincentz         | R. B. Wilson     |
| B. D. Smith      | R. Summers        | R. Vining           | B. Wingast       |
| D. E. Smith      | J. Sutton         | W. K. Viscusi       | S. Winter        |
| J. Smith         | J. Svejnär        | R. Vishny           | R. Wintrobe      |
| J. L. Smith      | L. E. O. Svensson | X. Vives            | A. Witte         |
| R. S. Smith      | C. E. Swan        | G. von Fustenberg   | D. Wittman       |
| S. Smith         | J. L. Swofford    | P. B. Voos          | B. Wolfe         |
| V. K. Smith      | G. Tabellini      | M. Waldman          | R. Wolff         |
| V. L. Smith      | P. Tandon         | D. G. Waldo         | M. Wolfson       |
| E. Smolensky     | V. Tanzi          | T. Wales            | K. I. Wolpin     |
| R. A. Snyder     | P. Taubman        | I. Walker           | K. Wong          |
| J. Sobel         | M. K. Taussig     | C. E. Walsh         | P. Wonnacott     |
| G. R. Solon      | L. Taylor         | H. Wan              | W. T. Woo        |
| J. Solow         | D. Teece          | P. Warr             | S. Woodward      |
| R. M. Solow      | L. Telser         | E. Wasensjoe        | G. Wright        |
| J. Sonstelie     | P. Temin          | M. W. Watson        | R. D. Wright     |
| R. H. Spady      | R. Thaler         | H. W. Watts         | M. Yaari         |
| B. Spencer       | S. E. Thiel       | R. N. Waud          | J. L. Yellen     |
| D. Spencer       | D. Thomas         | W. Weber            | J. M. Yinger     |
| M. Spicer        | J. Thursby        | P. Weil             | S. Yitzhaki      |
| M. Spiegel       | M. Thursby        | B. R. Weingast      | L. Young         |
| D. Spiegelman    | N. Tideman        | D. Weir             | G. A. Zarkin     |
| D. Spulber       | T. H. Tietenberg  | A. Weiss            | V. Zarnowitz     |
| T. N. Srinivasan | C. P. Timmer      | T. Weisskoff        | R. Zeckhauser    |
| J. E. R. Staddon | J. Tirole         | K. West             | S. P. Zeldes     |
| F. Stafford      | S. Titman         | F. Westfield        | D. Zilberman     |
| R. Staiger       | M. P. Todaro      | J. Whalley          | M. Zupan         |
| R. Startz        |                   |                     |                  |

ORLEY ASHENFELTER, *Editor*

## Report of the Editor

### *Journal of Economic Literature*

The *Journal of Economic Literature's* mission is to help members of the American Economic Association maintain a broad knowledge of economics in the face of powerful pressures toward specialization. This mission is effected by providing a guide to economics research and publications in the form of survey articles, book reviews, an annotated listing of new books, a compendium of tables of contents of current periodicals, a list of titles of recent Ph.D. dissertations, and selected abstracts of articles. The classified indexes of articles appearing quarterly in the *Journal* are brought together in an annual *Index of Economic Articles*.

The *Journal's* bibliographic departments consist of the Annotated Listing of New Books, the Contents of Current Periodicals, Selected Abstracts of Articles, and a List of Doctoral Dissertations in Economics. (This last department was taken over in 1987 by the *Journal* from the *American Economic Review*.) These departments are carried on in the *Journal's* Pittsburgh office under the direction of the Associate Editor, Drucilla Ekwurzel. Mary Kay Akerman serves as Assistant Editor and Professor Asatoshi Maeshiro of the University of Pittsburgh is an Editorial Consultant with general responsibility for the classification of articles in the *Journal's* Subject Index. The Pittsburgh office also benefits from the services of Pat Andrews, Liz Braunstein, and Beth Thornton. They have my thanks for their devoted and effective work throughout the year. Linda Scott stepped down from the position of Assistant Editor in May 1987 and I am grateful to her for her notable contributions to the *Journal*.

In 1987, the *Journal* provided annotations of over 1,250 new books, listings of approximately 1,300 issues of economics journals, and abstracts of over 2,400 articles. Also the 1981 and the 1984 *Index of Economic Articles* were published with the 1982

*Index* to follow shortly. The subject and author indexes of journal articles are available through computer access to the *Economic Literature Index* (ELI) of the DIALOG Information Retrieval Service. An article by Drucilla Ekwurzel and Bernard Saffran in the December 1985 issue of the *Journal* provides information about this service.

In recent years, because of the relentless increase in the number and size of economics journals, the *Journal's* bibliographic departments have grown substantially: whereas 1,348 pages were devoted to these departments in the 1980 issues of the *Journal*; in 1986 these departments took up 1,652 pages. This growth was held back in 1987 to 1,649 pages (or to 1,667 pages if the list of doctoral dissertations is included). In part, the containment in the expansion of these departments in 1987 resulted from implementing a rule by which no article would be catalogued under more than two 3-digit classifications of our Subject Index of Articles. This rule was introduced in our September issue and in a full year some 60 to 70 pages per year are likely to be saved by this measure. The opportunities for making more marginal changes of this sort are few and we have been looking into more radical measures to contain the size and cost of the *Journal*. This might mean the elimination of one of the *Journal's* bibliographic departments even though our surveys of members' opinions indicate that an important minority would regret this step. We would implement such a change with great reluctance, but given the financial constraints of the *Journal* we shall probably have to make an awkward decision of this sort in the next few years.

The Articles and Communications and Book Review departments of the *Journal* are managed from Stanford University. During 1987, the *Journal* published 9 major articles, 1 communication, and 166 book reviews. Moses Abramovitz helps me in reviewing

and monitoring the progress of our manuscripts and Alex Field supervises the Book Review department. We receive excellent support from Anita Makler, Ann Vollmer, and Toni Haskell. They all have my thanks for their valuable efforts. In addition, we have profited from the generous advice of our Board of Editors. Five members of the Board have completed their terms as of the end of 1987, namely, David Bradford, Stephen Lewis, Richard Marston, Thomas Mayer, and Robert Pollak. I am most grateful for their help. I am very pleased that Richard Marston, Thomas Mayer, and

Robert Pollak have agreed to serve another term on the Board. I shall be proposing to the AEA Executive Committee that Harvey Rosen of Princeton University and Mark Rosenzweig of the University of Minnesota join the Board next year. I thank all the members of the Board for their very real contributions of the *Journal*. Also, we have called upon a number of referees who have often made very substantial contributions to the quality of our articles. Their efforts are genuinely appreciated. We list below the referees used during 1986 and 1987.

H. Aaron	R. Fels	H. M. Levin	R. Ruggles
G. Ackley	S. Fischer	T. E. MaCurdy	P. A. Samuelson
G. A. Akerlof	C. Freeman	C. F. Manski	J. Scadding
J. Adams	R. T. Freeman	P. Marer	F. M. Scherer
R. Aliber	B. M. Friedman	R. C. O. Matthews	T. W. Schultz
J. G. Altonji	C. A. E. Goodhart	L. W. McKenzie	A. Sen
B. Arthur	C. Goodwin	R. I. McKinnon	S. Shavell
M. Baily	R. J. Gordon	G. M. Meier	T. Sicular
W. J. Baumol	G. Grossman	G. O. Mensch	F. A. Sloan
G. S. Becker	T. W. Guinnane	E. S. Mills	R. S. Smith
C. S. Bell	M. Gunderson	O. Mitchell	R. Solow
A. Bergson	J. Gurley	G. H. Moore	R. W. Staiger
A. Berry	R. W. Hansen	R. J. Murnane	D. A. Starrett
A. S. Blinder	J. Heckman	R. F. Muth	H. Stein
G. J. Borjas	R. Heilbroner	A. F. Nakamura	W. E. Steinmueller
J. A. Brander	E. A. Hewett	R. R. Nelson	W. F. Stolper
G. F. Break	J. Hirshleifer	P. Newman	P. Temin
E. Burmeister	C. A. Holt	J. R. Norsworthy	J. Tobin
G. Cain	H. J. Holzer	W. E. Oates	L. D. Tyson
C. S. Carson	P. W. Howitt	M. V. Pauly	D. Usher
B. R. Chiswick	R. P. Inman	J. A. Pechman	J. J. Van Duijn
G. C. Chow	B. Johnston	A. M. Polinsky	J. M. Vernon
A. Clark	D. W. Jorgenson	R. D. Portes	W. Vickrey
R. D. Cooter	F. T. Juster	P. M. Rappoport	E. R. Weintraub
R. Cornwall	A. C. Kelley	T. G. Rawski	D. Weir
J. C. Cox	J. W. Kendrick	M. Reich	A. Wildavsky
S. H. Danziger	J. F. Kennan	U. Reinhardt	R. Wilson
P. Danzon	S. R. King	L. G. Reynolds	D. Wittman
F. De Leeuw	J. Kornai	J. D. Richardson	P. Wonnacott
E. F. Denison	R. Kravis	J. G. Riley	P. Yotopoulos
P. A. Diamond	N. Lardy	M. Riordan	E. J. Zajac
R. A. Easterlin	L. J. Lau	M. Roemer	
S. Fabricant	S. Lebergott	H. S. Rosen	
H. S. Farber	R. D. Lee	M. Rothschild	

JOHN PENCANEL, *Editor*



## Report of the Editor

### *Journal of Economic Perspectives*

Several years of discussion and planning for the launching of a new Association journal finally reached fruition in August of this year with the publication of the first issue of the *Journal of Economic Perspectives*. The second issue was published in November, and the reception to both has been overwhelmingly positive.

The early part of the year entailed a number of time-consuming and difficult design and production decisions. The Editors wish to acknowledge the assistance of Laurel Cantor in the design of the *Journal*. She prepared numerous mock-ups of different typefaces, styles, and covers, and was very helpful in her dual role of providing guidance while being responsive to the tastes and choices of the Editors. Although we did not make full use of recent results in competitive bidding theory (an article concerning this theory will appear in the *Journal* shortly), we solicited bids from four printers and four compositors, and after extensive discussions with the Nashville office, agreed to use Banta Company and Science Typographers, Inc. These are the same companies presently used in producing the *American Economic Review*. The contract was awarded to them on the basis of their low bid and their previous excellent track record with the Association.

The editorial process, in which the Editors and Associate Editors solicit articles on different topics, has proven to work well. By drawing on a group of economists from different geographical areas and with different interests, it has enabled a diversity of interests and perspectives to be represented in the journal. While most of the Associate Editors have done the absolutely superlative job that was anticipated, some of them have been less active, a problem to which the Editor and Co-Editor have been attentive. Three Associate Editors have also found it necessary to resign under the press of other commitments.

Heading into 1988, the membership of the editorial board includes: Henry J. Aaron,

The Brookings Institution; Stanley Fischer, Massachusetts Institute of Technology; Dwight M. Jaffee, Princeton University; Edward P. Lazear, University of Chicago and The Hoover Institution; Mark J. Machina, University of California at San Diego; Charles F. Manski, University of Wisconsin; John E. Roemer, University of California at Davis; Kenneth S. Rogoff, University of Wisconsin; Bernard Saffran, Swarthmore College; Steven C. Salop, Georgetown University Law Center; Lawrence H. Summers, Harvard University; Hal R. Varian, University of Michigan; Janet L. Yellen, University of California at Berkeley.

We also continue to be extremely pleased with the responsiveness of those authors we have solicited to contribute to the *Journal*. We have received an almost universally enthusiastic response from our potential authors; they have not only agreed to write papers, but they have delivered them (almost) on time. They have put energy and care into writing articles which are responsive to the objectives of the *Journal*, and they have responded with patience and attention to the suggestions we have provided for the revision of their articles.

The first two issues of the *Journal* totaled 400 pages in length. They included twenty full-length articles, including symposia on tax reform, merger policy, and arbitrage, as well as individual papers on other topics. In addition, the issues included ten shorter introductions or features, including three regular features of the *Journal*: Puzzles, Recommendations for Further Reading, and Anomalies. The production process is already well-advanced for the next several issues. As of the time this report is being written, we have ten more full articles being corrected in the galley stage, along with eleven shorter introductions, features, or comments on longer articles. Comments have been completed on another eight papers, and we are expecting the final drafts of those papers in the office before the meetings. In

addition, we have another group of ten first drafts in the office for which we are currently producing comments.

Although agreements to write papers always contain an element of uncertainty, we are currently having no difficulty in attracting articles that will maintain the high quality of the first two issues.

The employees in our Princeton office deserve commendation for their excellent work. Carolyn Moseley played an invaluable role in the innumerable production decisions associated with the launching of a new journal, and she continues to manage the office, deal with the printer and compositor on a day-to-day basis, and help with proofreading articles.

The Editors feel that they cannot overstate the important role that Managing Editor Timothy Taylor has played in the new *Journal*. He has not only managed the *Journal* and kept it to some sort of schedule, but has ensured that the broader objectives of the journal have been met. He has performed the difficult task of persuading authors to amend and rewrite their articles with delicacy and skill, and has shown that it is possible to edit papers and increase their clarity and accessibility, while still retaining the distinctive voice of each author.

Also, the Editors would like to express their thanks to the many people at Princeton University and the Woodrow Wilson School of Public and International Affairs for their help. The *Journal* offices are located in the Wilson School, which has been a great help in gathering supplies, providing copying services, keeping track of expenses, putting in new phone lines and equipment, and smoothing out all the other bumps that come with setting up an office.

Two broad policy issues should be raised at this time for the consideration of the Executive Committee: Several individuals have written or spoken to the Editors to express their concern that the *Journal* is publishing primarily solicited papers, and to express their desire for a more open process of direct submissions. This question is a difficult one. We are most sympathetic to the concern and objective that the articles that appear in the *Journal* should reflect the broad

range and diversity of the economics profession; indeed, that is one of the *Journal's* reasons for being. There is certainly a tension between a policy of soliciting papers and an attempt to be open to the profession as a whole.

But at least for the present, several practical concerns dictate that the *Journal* continue to focus on solicited manuscripts. First, the *Journal* does not have the editorial resources with which to screen a large number of unsolicited submissions. (The *American Economic Review*, for example, receives eight manuscripts for every one it can publish.) The terms on which the Associate Editors accepted their jobs explicitly stated that they would be responsible for soliciting and editing manuscripts, not for refereeing papers. If outside referees were used, the Editors would have to review many of the manuscripts, including all manuscripts with recommendations that could be interpreted as favorable. As it is, the *Journal* is consuming more of their time than the Editors had anticipated.

Second, the *Journal* does not have much space for unsolicited articles. Much of each issue is taken up with organized symposia, in which a number of writers address the same topic, with regular features, or with other solicited articles. Given the present size of the *Journal*, we could only envisage printing a maximum of perhaps 8–10 unsolicited articles each year. Reviewing a great many manuscripts in order to publish so few does not appear to be a wise way in which to spend the *Journal's*, or the Association's, resources.

To sum up: The *Journal* will attempt to be responsive to suggestions about possible topics and potential authors from all sources. We welcome letters and proposals for various topics. When unsolicited articles arrive in our offices, we will certainly skim through them and respond to the authors. But at the present time, we recommend that the *Journal* remain primarily one of solicited articles. We would like the Executive Committee to express an opinion regarding this planned course of action.

A second issue on which the Executive Committee may wish to express themselves concerns the publication of the annual

Richard T. Ely Lecture and the Presidential Address. This year, the Ely Lecture will appear in the May issue (the *Papers and Proceedings* issue) of the *American Economic Review* and the Presidential Address will appear in the March issue of the *AER*, as has been traditional practice. Our suggestion is that next year's Ely Lecture be published in the *JEP*, and that the Presidential Address be published in the *JEP* the year after. President-Elect Joseph Pechman is enthusiastic about this suggestion. However, we believe that these lectures should not oscillate back and forth between different journals, but should have a permanent home. The Association should therefore make a long-run commitment concerning the place of publication. The proposed arrangement

has the advantage of leaving the *American Economic Review* as the refereed research journal of the Association.

Requested Actions:

1. Approval of new members of the Editorial Board: Dwight M. Jaffee, Princeton University; John E. Roemer, University of California at Davis; Kenneth S. Rogoff, University of Wisconsin.
2. Approval of journal policy with regard to unsolicited manuscripts.
3. Decision concerning place of publication of Richard T. Ely Lecture and Presidential Address.

JOSEPH STIGLITZ, *Editor*

CARL SHAPIRO, *Co-Editor*

# Report of the Director

## Job Openings for Economists

The total number of new jobs listed this year increased significantly, reaching a record high of 1,924. This is almost a 25 percent increase over the number (1,561) listed last year and is the first time since 1984 that new jobs listed has increased. Both academic and nonacademic listings increased; the former from 1,134 in 1986 to

1,294 in 1987, the latter from 427 to 630. Table 1 shows total listings (employers), total jobs, new listings, and new jobs by type (academic and nonacademic) for each issue of *JOE* in 1987.

Table 2 shows the number of employers by category (four-year colleges, universities with graduate programs, federal government,

TABLE 1—JOB LISTINGS FOR 1987

Issue	Total Listings	Total Jobs	New Listings	New Jobs
Academic				
February	65	113	53	90
April	50	93	47	84
June	37	59	34	76
August	46	99	41	89
October	177	442	163	421
November	146	335	146	335
December	207	510	33	199
Subtotal	728	1,651	557	1,294
Nonacademic				
February	18	65	15	44
April	22	84	21	79
June	30	71	28	61
August	24	83	19	66
October	37	148	31	134
November	38	122	38	122
December	62	229	31	124
Subtotal	231	802	183	630
Total	959	2,453	750	1,924

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1987

Issue	Four-Year Colleges	Universities with Graduate Programs	Federal Government	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	20	45	7	—	1	—	8	2	83
April	17	33	8	1	5	—	6	2	72
June	12	25	9	1	6	5	8	1	67
August	11	35	8	2	6	—	5	3	70
October	52	125	13	3	9	1	8	3	214
November	57	89	13	3	5	2	15	—	184
December	91	116	23	3	12	1	22	1	269
Total	260	468	81	13	44	9	72	12	959

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1987

Fields <sup>a</sup>	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	57	56	39	49	215	286	278	880
Growth and Development (100)	21	31	20	19	50	37	57	235
Econometrics and Statistics (200)	27	25	21	12	78	54	101	318
Monetary and Fiscal (300)	21	22	21	21	97	75	131	388
International Economics (400)	19	14	15	17	68	62	93	288
Business Administration, Finance, Marketing and Accounting (500)	15	15	14	13	43	37	61	198
Industrial Organization (600)	27	29	19	16	54	47	83	275
Agriculture and Natural Resources (700)	13	10	9	13	26	18	31	120
Labor (800)	13	12	11	13	42	40	70	201
Welfare and Urban (900)	13	8	8	16	51	40	67	203
Related Disciplines (A00)	7	4	2	2	11	5	11	42
Administrative Positions (B00)	6	6	6	8	12	5	11	54
Total	239	232	185	199	747	606	994	3,202

etc.) for each of the 1987 issues. Academic institutions continue to be the major source of jobs, about 75 percent of the total number of employers listing vacancies. The pattern this year is roughly the same as it has been for several years.

Table 3 shows the number of citations by field of specialization. General economics (000) led, followed by monetary and fiscal (300) and econometrics and statistics (200).

This continues the pattern that has prevailed for several years.

Violet Sikes is responsible for everything to do with the publication and distribution of *JOE*. She continues to do excellent work and I am very grateful for her dedication and loyalty.

C. ELTON HINSHAW, *Director*

## Report of the Committee on Economic Education

The Committee continued its varied efforts to enhance economics instruction at the college level and to facilitate research on instruction in economics. The Committee also worked closely throughout the year with the Joint Council on Economic Education in promoting improved economics instruction at all levels of education. The Committee typically meets once each year, and it carries on its activities largely through informal interactions between the Chair and the various members. Its works with the JCEE is accomplished in several ways: the Chair serves on the Executive Committee of the JCEE Board of Trustees, and the Chair works closely with the JCEE President and JCEE Director of Research on matters of mutual interest.

Important progress was made this year in initiating a new program of research and research training in economic education. This effort is funded by a generous three-year grant to the JCEE from the Pew Trust. The program has two major components. One is to develop through special surveys a body of national baseline data on economics instruction at the 12th grade level that will open up new areas of research. A preliminary report on the results of these surveys was presented by William J. Baumol and Robert J. Highsmith at the December AEA meetings. (In due time, the survey data base will be released for use by other researchers; an announcement will appear in a future issue of the *Journal of Economic Education*.) The other is to attract more younger economists into research on economic education, much as was done in the late 1960's and early 1970's through a similar program underwritten by the General Electric Foundation. This is being accomplished now through the organization of two one-week summer seminars designed to acquaint participants with the literature in economic education research, the variety of research techniques that can be utilized, and the special problems associated with such research. The first seminar was held at Princeton University last August; a second seminar will be held in

summer 1988. Over 80 economists applied for the 20 openings in the 1987 seminar. The program did not end with the seminar, however. During the current academic year, participants, aided by small research grants, are conducting research projects that utilize the national data base. The participants will be reconvened this coming summer to discuss their progress and findings. The Committee hopes to schedule several of the resulting papers for the Economic Education session at the December 1988 annual meeting. Robert J. Highsmith of the JCEE is to be commended for his energy and efficiency in bringing this program into existence.

The Committee also cosponsored with the *Journal of Economic Education (JEE)* a conference in September 1987 on the state of economics principles textbooks. The conference, held at Indiana University-Purdue University at Indianapolis, brought together 50 participants, including 6 textbook authors, a number of economists, and a group of economic educators. The purpose was to discuss the strengths and weaknesses of the major textbooks, and what might be done to improve the quality of these textbooks. The conference was lively and featured major papers by Kenneth Boulding, Joseph Stiglitz, Carolyn Shaw Bell, and Michael Boskin. The consensus seemed to be that the textbooks are doing a reasonably good job of meeting their goals. More details about the conference can be found by reading the papers and accompanying responses by textbook authors which will be published in the Spring 1988 issue of the *JEE*. The conference was conceived by Kalman Goldberg; Dennis Weidenaar of Purdue University and Robin Bartlett of Denison University organized this conference. The Committee is much indebted to them for their initiative.

The Committee reviewed the condition of the *JEE*, based on the annual report of the *JEE* prepared by its editor, Kalman Goldberg. He reported that the volume and quality of the submissions continues to increase, with the backlog of accepted papers expand-

ing. There was discussion about whether to expand the size of the *JEE* in light of the flow of submissions and the practice of devoting one issue each year to the proceedings of special conferences, such as that on textbooks (see above) and another on the scope of economics which appeared in the Spring 1987 issue of the *JEE*. The editor expressed the hope that more members of the profession would subscribe to the *JEE* whose price is \$23 per year for its four issues. The *JEE* should be of particular interest to faculty members whose principal activity is teaching and also to those who are interested in research on the effectiveness of economics instruction at the college level.

The Committee's efforts to update its Teacher Training Program (TTP) continue. Arrangements have been worked out by William Walstad and Phillip Saunders with McGraw-Hill to publish a book in 1989 that will combine some of the materials from the *TTP Resource Manual* with other commissioned papers on economics instruction at the principles level. This volume will not replace the *TTP Resource Manual*, but will contain those portions of it that are most pertinent to instruction in the elementary course. In light of this development, the Committee hopes to launch a new approach to secure funding to develop a self-teaching approach, along with appropriate videotapes, so that these materials can be used by individual faculty members and teaching assistants who want to improve their teaching. Michael Salemi and Bruce Dalgaard will be reporting back to the Committee with their recommendations.

Little progress has been made on the problems that arise as a result of the growing numbers of foreign students who are serving as teaching assistant. The Committee discussed the possibilities of a survey of departments to learn more about the magnitude of the problem and what methods departments have developed to cope with these problems. A plan of action is being prepared by W. Lee Hansen that the Committee will consider in the near future.

The Committee also heard a report from Steven Buckles on the progress of the Ad-

vanced Placement Program in Economics which is being developed in collaboration with the College Board, The Educational Testing Service, and the JCEE; he chairs the committee appointed by the College Board to develop the economics examination. The AP Program permits departments to grant college-level credit to students who as high school students demonstrated through a special examination a knowledge of introductory economics. With the first testing set for late in the 1988-89 academic year, much is happening. Several pilot exams have been developed and pretested, and are now being reviewed by economists. An AP instructional package is being prepared for use by high school teachers in the AP courses. Finally, plans are underway to mount several regional workshops that will prepare AP teachers for the first round of AP course offerings. Again this year, the JCEE sponsored a breakfast at the AEA meetings to inform members of the economics profession about developments in the AP Program in Economics.

Finally, the Chair constituted a subcommittee to report by mid-1988 on the mission and future activities of the Committee on Economic Education. This step is motivated by a sense that the Committee should broaden its role, by increasing its efforts to inform, influence, and involve the profession in promoting improve economics instruction not only in the colleges but also in the elementary and secondary schools. Economics instruction is now mandated in at least 27 states, and economics instruction takes place in many schools and school districts. It is important that such instruction be improved because this is the only formal exposure to economics instruction for the many K-12 students who do not attend college and for those who do attend college but never take an economics course. This subcommittee will offer recommendations about how to undertake this task; it welcomes suggestions and comments which should be directed to the Chair of the Committee.

W. LEE HANSEN, *Chair*

## Report of the Committee on the Status of Women in the Economics Profession

The Committee on the Status of Women in the Economics Profession (CSWEP) was extremely active in 1987. In addition to arranging technical sessions and social events at the annual and regional meetings of the economics associations, CSWEP updated and produced *Women in Economics*, a roster of women economists containing information such as employer, educational background, fields of specialization, and number of publications. Copies were sent to the chairs of economics departments that grant Ph.D.s for use in filling faculty positions, as well as to all CSWEP members. Many thanks are due to Joan Haworth, the Committee's Membership Secretary, and her staff for completing this demanding task on time, updating CSWEP's mailing list throughout the year, and preparing special tabulations of the roster for employers who requested them.

Another major activity was to publish three issues of the CSWEP *Newsletter*. This year, the *Newsletter* has continued to focus on helping younger faculty members advance their careers, with articles on topics such as searching for senior academic jobs and surviving the tenure process. Each issue also contained a description of a particular economist's current job or career path, a book review, and a listing of job openings. The Committee thanks Katharine Lyall, who is now arranging for articles to be written, and Toni Foxx, who is responsible for the *Newsletter*'s production, for the excellent jobs they are doing.

Following a presentation by Belle Sawhill (CSWEP's former Chair) on double-blind reviewing to the AEA's Executive Committee, Orley Ashenfelter (Editor of the *American Economic Review*) proposed examining the effect of single- vs. double-blind reviewing procedures using manuscripts submitted to the *AER*. The evaluation is being conducted by Rebecca Blank of Princeton University, with considerable cooperation from the *AER*'s staff. The Committee is encour-

aged that more information on this topic is being gathered, although we continue to advocate the adoption of double-blind reviewing as a matter of principle—primarily because it is fairer for all groups against whom discrimination may exist, such as economists at less prestigious institutions or women.

The project to examine differences in the career paths of men and women with Ph.D.s in economics, which is being conducted by Sue Berryman and Arthur Kennickell and funded by the Russell Sage Foundation, has made little progress this year because of a lack of access to confidential data that are maintained by the National Academy of Sciences. These difficulties have recently been resolved and empirical results should be available in 1988.

Two new projects were begun this year. To facilitate employers' use of the roster of women economists, Judy Lave will work with Joan Haworth and her staff to prepare listings of women researchers by field and years of experience. The appropriate listings will automatically be sent to employers submitting job announcements for the CSWEP *Newsletter*. To keep the information current, the data for the roster will be updated each year (using the AEA's mailing list and questionnaires sent to those already on the roster); we will continue to produce "hard" copies of the roster only every other year.

The second project stems from a suggestion by Alice Rivlin to examine the process by which sessions and papers are chosen for the AEA's annual meeting. In the past, presidents-elect have used somewhat different approaches, including various ways of encouraging participation on the program by broader groups of economists. Recently, both Robert Eisner and Joseph Pechman have been particularly supportive of CSWEP's goals and have expanded the Committee's responsibilities for arranging sessions. But do the characteristics of participants in the final program or of the authors represented in the *Papers and Proceedings* depend on the



way the meeting was organized? For example, does using a program committee matter? We will examine these questions using data about recent annual meetings.

Finally, the Committee thanks Belle Sawhill, who completed her three-year term as Chair this year, for her extensive contributions. For example, she provided the impetus for a serious examination of the effects of single- vs. double-blind reviewing. She also initiated and obtained funding for the proj-

ect that is comparing career paths of men and women economists. Karen Davis, whose term also expired this year, contributed much as well. In particular, she took major responsibility for reviewing the papers presented at CSWEP-organized sessions at the last two annual meetings to determine which would appear in the *Papers and Proceedings*.

NANCY M. GORDON, *Chair*

## Report of the Committee on U.S.-Soviet Exchanges

The tenth U.S.-Soviet Economic Symposium was held in Tbilisi, September 1-4, at the Georgian Academy of Science's Institute of Law and Economics. The U.S. delegation consisted of ten economists: myself; six experts on U.S. agriculture who presented papers; three with expertise also on Soviet agriculture who served as discussants of the Soviet papers, one of whom also presented a paper. The Soviet delegation also consisted of ten persons, eight of whom presented papers on Soviet agriculture.

Most of the U.S. and Soviet papers complimented each other quite well. At the first session, two Soviet papers and one U.S. paper dealt with the adequacy, utilization, and preservation of the agricultural resource base. One paper from each delegation dealt with econometric and planning models. As Soviet methods of managing agriculture may shift under Gorbachev from administrative to economic levers (use of prices, rents, taxes, etc.), Soviet problems and methods could begin to converge with those of the United States. Another American paper discussed the effects of economic growth on agriculture while a Soviet paper described the effects of economic growth on the level of nonagricultural activities in rural areas. Another pair of papers dealt with agricultural policy reforms in both nations.

Two U.S. papers had no Soviet counterpart, one on U.S. farm debt and one

comparing Soviet and U.S. experiences in agricultural pricing. Both papers evoked considerable discussion. Finally, a triad of Soviet papers dealt with the peculiarities of the policies and structure of the Georgian agro-industrial complex. These were of considerable interest in terms of recent agricultural reforms because Georgia (and Latvia) are leaders in experimentation.

From the standpoint of the American participants, the Symposium was a success. The discussions were interesting and quite frank (clearly reflecting *glasnost*) and chairpersons had difficulty keeping discussants to the formal time limits. The Georgian Academy of Sciences had asked for the conference to be held in their republic and this was reflected in the warmth of their hospitality which would be difficult to surpass. Tbilisi is a lovely old city in the foothills of the Caucasus and a fine site for a conference. While in Tbilisi, we did a fair amount of sightseeing; visiting a farm, a factory, and a fantastic collective farm market; and experiencing interesting cuisine and excellent wine. In addition to our stay in Tbilisi, we had a few days of sightseeing in Moscow.

The next symposium is scheduled to be held in the United States sometime in 1988. We have not yet settled on a topic for discussion.

FRANKLYN D. HOLZMAN, *Chair*

## Report of the Committee on U.S.-China Exchanges in Economics

Exchanges in economics with the People's Republic of China continued to expand in 1987. Besides the numerous individual arrangements made by scholars and graduate students from each country to visit the other, important official exchanges took place.

The U.S. Committee on Economics Education and Research in China, in cooperation with the Chinese Committee on Economic Exchanges with the United States and with financial support from the Ford Foundation, has been responsible for the following activities.

First, a year-round graduate economics training program at the People's University continued its operations for the third year, with T. Dudley Wallace, Duke University, and Donald D. Hester, University of Wisconsin, teaching the fall of 1987, and Luis Guasch, University of California at San Diego, and Nicholas H. Stern, London School of Economics, teaching in the spring of 1988. A new graduate training program was established at Fudan University with Daniel Suits, Michigan State University, and Kar-Yiu Wong, University of Washington at Seattle, teaching in the fall of 1987, and Andrew Feltenstein and Anne Sibert, both of the University of Kansas, teaching in the spring of 1988. Anyone interested in teaching in one of these two graduate programs should write to Dr. Todd Johnson, Executive Director of the Committee, CSCPRC, National Academy of Sciences, 2101 Constitution Avenue, N.W., Washington, D.C. 20418.

Second, graduate students, sponsored by the Chinese State Education Commission, continued to come to the United States and Canada, with twenty-seven arriving in the fall of 1987 and twenty-one having been selected in late fall 1987 to apply for admission in September 1988.

Third, a summer workshop on Information and Coordination in the Enterprise and the Economy, organized by Lawrence J. Lau of Stanford University, took place at the People's University in June-July 1987, serv-

ing the students of the year-around graduate economics training program and research economists from the government. Other lecturers of the workshop include Michael Riordan, Hoover Institute at Stanford; Peter Hammond and B. Douglas Bernheim, Stanford University; and Debraj Ray, Indian Statistical Institute in New Delhi.

In September, the Chinese Committee on Economic Exchanges with the United States visited the United States and held a joint meeting with the U.S. Committee on Economics Education and Research in China. The former committee, headed by Vice President Huang Da of The People's University, consists of officials of the Chinese State Commission of Education responsible for economics education as well as representatives of seven major universities, including Beijing University, Fudan University, Jilin University, Nankai University, the People's University of China, Wuhan University, and Xiamen University. The two committees decided to continue the above three activities for another three years, until 1990. The Chinese Economics Committee plans to select candidates to be visiting scholars in the United States for a period of up to one year. The U.S. Committee is also exploring possible joint research by European and American scholars with the staff of the Chinese Academy of Social Sciences and the Institute of Economic System Reform of the Chinese State Council.

A Conference on PRC-U.S. Economic Cooperation was held from June 28 to July 2 in Wuhan. It was sponsored by the American Committee on Asian Economic Studies and was organized by Pei-kang Chang and Shao-kung Lin, Huazhong University of Science and Technology, and M. Jan Dutta, Rutgers University. Numerous papers were presented by American and Chinese scholars.

A delegation of economists representing the World Bank and including AEA members Carl Christ, Peter Diamond, Gustav

Ranis, and Jacob Siegel visited Shanghai in August to give lectures and to exchange views with the Chinese Review Committee, appointed by the Chinese State Education Commission and headed by President Xie Xide of Fudan University, on economics curriculum reform at Chinese universities. They recommended the inclusion of microeconomics, macroeconomics, public finance, monetary economics, and accounting as subjects for the curriculum, which, until now, has consisted mainly of Marxian economics.

Ma Hong, Head of the Research Center of Economic, Technological, and Social Development of the State Council, led a delegation to visit the United States in November to widen contact with American institutions engaged in economic research relevant for policy analysis.

GREGORY C. CHOW, *Chair*

## Report of the Representative to the National Bureau of Economic Research

The National Bureau of Economic Research studies a wide variety of economic issues. Much of this research is conducted by economists working individually, but a large part is organized into special projects. To disseminate the results of this research, during 1987 the NBER issued over 350 working papers, published 7 books and 2 annual journals, and circulated the monthly *Digest* and the quarterly *Reporter*. In addition, the NBER held numerous conferences and workshops, including the 6-week Summer Institute. Finally, the NBER awards six Olin Fellowships each year to allow young economists to spend a year conducting empirical research on topics of their own choosing.

*Projects.* NBER's projects generally bring a dozen or more researchers together to work on a common topic. The projects' findings are usually distributed initially as NBER working papers, and final versions are then published as Bureau books. During 1987 the following projects (organizer in parenthesis) were underway: Economics of Aging (David Wise), International Economic Coordination (Martin Feldstein), Developing Country Debt (Jeffrey Sachs), Exchange Rate Misalignment (Richard Marston), Political Economy (Robert Baldwin), State and Local Government Finance (Harvey Rosen), U.S. in the World Economy (Martin Feldstein), International Migration (Richard Freeman), Macroeconomics (Stanley Fischer), Mergers and Acquisitions (Alan Auerbach), Strategic Trade Policy (Paul Krugman), Tax Policy and the Economy (Lawrence Summers), Trade Relations (Robert Baldwin and David Richardson).

*Programs.* NBER's ongoing programs generally meet twice during the academic year, and once, for a longer period, during the Summer Institute. Bureau programs (with directors in parentheses) are Economic Fluctuations (Robert Hall), Financial Markets and Monetary Economics (Benjamin Friedman), International Studies (William

Branson), Labor Studies (Richard Freeman), Taxation (David Bradford), Development of the American Economy (Robert Fogel), Health Economics (Victor Fuchs and Michael Grossman), and Productivity (Zvi Griliches).

*Other Conferences.* In addition to the conferences and workshops conducted as part of NBER projects and programs, the NBER also holds larger conferences open to individuals in government and academe. During 1987 these included a session of the Conference on Income and Wealth on the Measurement of Saving, Investment, and Wealth, organized by Robert Lipsey and Helen Tice; a Universities Research Conference on Labor Markets and the Macro Economy, organized by John Abowd; and a second URC on Risk in Financial Markets, organized by Vance Roley. Procedures and deadlines for submitting papers to these conferences are posted in economics departments and announced in the NBER *Reporter*.

*Summer Institute.* During the summer of 1987, 435 economists from 124 different universities and research organizations attended the Summer Institute, including 154 participants who came for the first time. Participants stayed for periods ranging from 4 days to 2 months. There were 237 papers presented in 29 separate sessions. A list of these papers is available from the NBER on request.

*Working Papers.* NBER working papers are circulated to economics department libraries around the world, and are available for \$2 per title from the NBER. Subscriptions to the entire series are available for \$300 per year inside the United States and Canada, \$500 other foreign.

*Periodicals.* The monthly NBER *Digest* provides brief summaries of working papers of general interest. The *Reporter* contains longer summaries of recent research, abstracts of all working papers, announcements of the publication of NBER books, a

survey of economic forecasters, and a calendar of forthcoming conferences. Both these periodicals are available on request.

*Olin Fellows.* The six Olin Fellows for the 1987-88 academic year are Mark Bills, Alberto Giovannini, Glenn Hubbard, N. Gregory Mankiw, Peter Reiss, and Christina Romer. These fellowships are open to young empirical economists at any academic institution. Application deadlines are announced in the *Reporter*.

During 1987, Martin Feldstein continued as President of the NBER and Geoffrey Carliner continued as Executive Director. Further information on NBER activities is available in the *Reporter* or from Geoffrey Carliner, NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138.

DAVID KENDRICK, *Representative*

## Report of the Representative to the American Association for the Advancement of Science

The American Association for the Advancement of Science is a federation of scientific organizations, as well as an association of over 135,000 individual members. Its objectives are "to further the work of scientists, to facilitate cooperation among them, to foster scientific freedom and responsibility, to improve the effectiveness of science in the promotion of human welfare, and to increase public understanding and appreciation of the importance and promise of the methods of science in human progress."

The discipline of Economics is grouped with Political Science, Sociology, Geography, and some related social sciences in Section K, one of the twenty-one divisions within the AAAS. Most economists list Section K as their primary affiliation, and some hold dual or multiple membership in other sections. Overall, the four social science sections comprise just under 10 percent of the AAAS-affiliated membership.

The AAAS offers economists a unique opportunity to interact with the scientific community at large. This includes the following channels: the AAAS Annual Meeting, joint sessions with other organizations at their professional meetings, specialty conferences, *Science*, *Science* 87, and AAAS-sponsored books. In addition, the organization is active through the work of its offices and standing committees on Science and Technology Education, Public Sector Programs, Opportunities in Science, International Science, Scientific Freedom and Responsibility, and Science, Arms Control and National Security.

Two sessions at the 1987 AAAS Annual Meeting, held in Chicago in February, were devoted to economics. The symposium entitled "Technology and Economic Change" included Wassily Leontief as chair, and papers by Richard Cyert and David Mowery, Faye Duchin, Holland Hunter, Nathan Rosenberg, and Susan Sanderson. The symposium on the subject of U.S.-Japanese

Trade included noted authorities from both countries.

A joint AEA/AAAS session was held at the 1987 Allied Social Science Meetings on "The Competitiveness of Natural Resource Industries." The session was chaired by David Gulley and included papers by Carol Taylor, Alberta Charney and Thomas Oxford, Richard Newcomb and Stanley Reynolds, Adam Rose, C. Y. Chen and Shih-Mo Lin, and Frank Wolak and Charles Kolstad.

Economics representation in the publication policies of *Science* was enhanced during 1987 with the appointment of Elizabeth Bailey to its Editorial Board. For the second year in a row, several feature articles were written by economists, including George Borjas and Marta Tienda, Vernon Briggs, Gregory Chow, Robert Eisner, Henry Farber, Benjamin Friedman, Zvi Griliches, Lester Lave, Mark Machina, and John Williamson.

At the beginning of 1987, twenty-five economists were fellows of the AAAS. These ranks were increased by two during the year with the appointment of Robert Evenson, Walter Isard, and Mancur Olson as new fellows.

In order to provide a broader based and more active AEA representation to the AAAS, it was decided to expand this function to a Liaison Committee. The committee will be composed of 6 members: the current AEA representative, a member of the Editorial Board or Board of Reviewing Editors of *Science*, AEA members with backgrounds in the physical, biological, behavioral and engineering sciences, and the past AEA representative. Accordingly, the 1988 Liaison Committee will consist of Adam Rose, Robert Solow, Gardner Brown, Vernon Smith, Faye Duchin, and Roger Bolton.

Further information about the AAAS is available from Marge White, 1333 H Street, N.W., Washington, D.C. 20005.

ADAM ROSE, *Representative*

## Temporary Equilibrium

### Selected Readings

Edited by Jean-Michel Grandmont

The collection of journal articles reproduced in this volume provide a synthesis of the progress made in the last fifteen years in the theory of temporary equilibrium. This is the first comprehensive collection of readings on the topic and is an essential tool for any professional economist or graduate student interested in the long-standing issue of providing macroeconomic theory with adequate microeconomic foundations.

April 1988, 512 pp.

\$49.95 (tentative) Casebound/ISBN 0-12-295145-X

\$19.95 (tentative) Paperback/ISBN 0-12-295146-8

## Law and Economics

### An Introductory Analysis, Second Edition

Werner Z. Hirsch

A variety of important legal topics are subjected to economic analysis by drawing on microeconomic theory and presenting a variety of empirical results. The second edition has undergone major revisions since the first edition was completed in 1978.

May 1988, 464 pp. (tentative) \$35.00 (tentative)

Casebound/ISBN 0-12-349481-8

## Security Markets

### Stochastic Models

Darrell Duffie

These theories of security markets deal principally with the allocational role and valuation of financial securities in a competitive setting. They provide a unified general equilibrium framework for recent advances in finance.

April 1988, 464 pp. (tentative) \$39.50 (tentative)

Casebound/ISBN 0-12-223345-X



**ACADEMIC PRESS**

Harcourt Brace Jovanovich, Publishers  
San Diego, CA 92101-4311

San Diego  
41058

New York

Boston

London

Sydney

Tokyo

Toronto

Credit card orders call toll free 1-800-321-5068.  
From Missouri, Hawaii, or Alaska 1-314-528-8110.  
Prices are in U.S. Dollars and are subject to change.



# INDEX OF ECONOMIC ARTICLES

prepared by

*The Journal of Economic Literature*  
of the  
*American Economic Association*

- ✓ Each volume in the **Index** lists articles in major economic journals and in collective volumes published during a specific year.
- ✓ No other single reference source covers as many articles classified in economic categories as the **Index**.

**Index** volumes XI-XXII covering 1969-1980 are available at \$60.00 each.

The following two part **Index** volumes are now ready for delivery at \$90.00 per set:

Volume	Year
XXIII	1981
XXIV	1982
XXV	1983
XXVI	1984

Note: AEA members  
are entitled to a  
30% discount.

*an  
indispensable  
tool for...*

ECONOMISTS  
REFERENCE LIBRARIANS  
RESEARCHERS  
TEACHERS  
STUDENTS  
AUTHORS

Payment required in advance. Prices include shipping charges; allow 4-6 weeks for delivery. Please send your check or money order (net of applicable discount) payable in United States dollars drawn on a United States bank to:

American Economic Assn. - **Index**  
P.O. Box 20111  
Nashville, TN 37202

Address inquiries or other correspondence to:  
Journal of Economic Literature, P.O. Box 7320  
Oakland Station, Pittsburgh, PA 15213.

# THE ECONOMICS INSTITUTE

Gateway to Successful  
Master's and Doctoral Degree Studies  
in Economics, Agricultural Economics, Business and Administration

*Since its establishment in 1958 under the sponsorship of the American Economic Association, the Institute has provided specialized preparatory training and orientation to over 6,000 students from over 120 countries en route to over 320 universities in the United States and other English-speaking countries.*

## PROGRAM FEATURES

- Six to forty-six week individualized programs.
- Intensive and comprehensive review or supplementary preparatory training in:  
**English, Computer Usage, Economic Theory, Mathematics, Statistics, Accounting, Management, Finance and Marketing.**
- Standardized test preparation (TOEFL, GRE, GMAT).
- Orientation to U.S. campus and community life.
- University placement assistance services.
- Certificate and Diploma Programs for short-term professional trainees.

## END RESULTS for

- **Associated universities:** Improved admissions procedures and a source of additional high-quality foreign students.
- **Foreign students:** Greater accessibility to U.S. universities.
- **Sponsoring organizations:** Reduced total costs.
- **And for all three:** Better grade performance records in degree programs and more rapid completion of degree requirements.

## ORGANIZATION

### **AEA Policy and Advisory Board:**

Edwin S. Mills, Northwestern University, Chairman

Lance E. Davis, California Institute of Technology

Koichi Hamada, Yale University

Joseph Havlicek, Ohio State University

Teh-wei Hu, University of California, Berkeley

Samuel A. Morley, Vanderbilt University

Ray Marshall, University of Texas at Austin

Stefan H. Robock, Columbia University

### **Director**

Wyn F. Owen, University of Colorado

**The Economics Institute, 1030 13th Street, Boulder, Colorado, 80302, USA;  
Telephone: (303)492-3000; Telex: 450385 ECONINST BDR; FAX: (303)492-3006**

# Supply-Side Tax Policy: Its Relevance to Developing Countries

by Ved P. Gandhi, Liam P. Ebrill, George A. Mackenzie, Luis A. Manas-Anton, Jitendra R. Modi, Somchai Richupan, Fernando Sanchez-Ugarte, and Parthasarathi Shome

Are the tax policy prescriptions of the supply-side approach applicable to developing countries, given their economic and institutional structures and economic policy environment? Are there other tax policy prescriptions and tax reforms that are supportive of the growth objectives of these countries? The twelve articles included in this volume examine selected aspects of these questions, both in theory and in practice.

Price: \$20.00 (388 pages; ISBN 0-939934-91-4)

Available from: Publication Services • Box No. E-387

International Monetary Fund • Washington, D.C. 20431 • U.S.A.

Tel. (202) 623-7430 • Cable Address: Interfund

## JOB OPENINGS FOR ECONOMISTS

Available only to AEA members and institutions that agree to list their openings.

### Annual Subscription Rates

U.S.A., Canada, and Mexico (first class): \$15.00, regular AEA members and institutions  
\$ 7.50, junior members of AEA  
All other countries (air mail): \$22.50, regular AEA members and institutions  
\$15.00, junior members of AEA

Please begin my issues with:

☐ February ☐ April ☐ June ☐ August ☐ October ☐ December

Name \_\_\_\_\_

First

Middle

Last

Address \_\_\_\_\_

City

State/Country

Zip/Postal Code

Check one:

- ☐ I am a member of the American Economic Association.  
☐ I would like to become a member. My application and payment are enclosed.  
☐ (For institutions) We agree to list our vacancies in JOE.

Send payment (U.S. currency only) to:

THE AMERICAN ECONOMIC ASSOCIATION  
1313 21st Avenue South  
Nashville, Tennessee 37212

# ANALYSIS & INSIGHTS

## Studies in economics from Chicago

### THE UNITED STATES IN THE WORLD ECONOMY

Edited by MARTIN FELDSTEIN

"This is an extraordinarily comprehensive and valuable collection that deserves a wide readership. *The United States in the World Economy* will appeal to specialists in both macroeconomics and international economics as a valuable reference work. The lucidity of the papers makes the book accessible to a much wider audience as well, those readers who are economically literate though not necessarily professional economists."

—Stanley Fischer, Massachusetts Institute of Technology

Paper \$24.95 712 pages 40 line drawings

Library cloth edition \$74.95

An NBER Conference Report

Now in paper

### TRADE AND EMPLOYMENT IN DEVELOPING COUNTRIES

Volume 3: Synthesis and Conclusions

ANNE O. KRUEGER

"This volume presents evidence which is critical to an understanding of the potential benefits of liberalization in LDCs. A familiarity with its conclusions . . . is essential for those concerned with international trade and economic development."—Edward Tower, *Southern Economic Journal*

Paper \$12.95 229 pages

An NBER Project Report

### EXCHANGE RATE THEORY AND PRACTICE

Edited by JOHN F. O. BILSON

and RICHARD C. MARSTON

"This volume is going to become a staple in any well-stocked international finance scholar's office."—Michael Melvin, *Journal of Economic Literature*

Paper \$16.95 538 pages

An NBER Conference Report

### MERGERS AND ACQUISITIONS

Edited by ALAN J. AUERBACH

*Mergers and Acquisitions* surveys, in a nontechnical format, some of the important issues arising from the recent takeover boom, offering a valuable resource for making appropriate policy decisions about the costs and benefits of corporate takeovers.

Cloth \$17.95 120 pages 8 line drawings

An NBER Project Report

### INTERNATIONAL ASPECTS OF FISCAL POLICIES

Edited by JACOB A. FRENKEL

The research reported in this volume is a timely analysis of current thinking on determinants of international fiscal policy, effects of shifts in fiscal policy, and the expectations concerning future policy change.

Cloth \$48.50 384 pages 38 line drawings

An NBER Conference Report

### PENSIONS IN THE U.S. ECONOMY

Edited by ZVI BODIE,

JOHN B. SHOVEN, and DAVID A. WISE

*Pensions in the U.S. Economy* reports on retirement saving of individuals and the saving that results from corporate funding of pension plans as well as on aspects of the plans themselves from the employee's point of view.

Cloth \$28.00 208 pages 18 line drawings

An NBER Project Report

### ECONOMIC BEHAVIOUR IN ADVERSITY

JACK HIRSHLEIFER

*Economic Behaviour in Adversity* brings together ten important essays, including several previously unpublished historical investigations and innovative theoretical analyses, on how individuals and societies function in times of disaster and conflict.

Cloth \$35.00 320 pages 22 line drawings

**THE UNIVERSITY OF CHICAGO PRESS**

5801 South Ellis Avenue, Chicago, IL 60637

*150 Years of Publishing Tradition . . .*

## **Allen & Unwin, Inc. is now UNWIN HYMAN, INC.**

In 1986 the venerable British publishing houses of George Allen & Unwin and Bell & Hyman merged to form Unwin Hyman, one of the largest independent British publishers. The U.S. division of Unwin Hyman, formerly Allen & Unwin, Inc., will now be called Unwin Hyman, Inc. and will continue its tradition of, and commitment to, publishing the finest in scholarly and general interest titles.

### ***Landmarks***

- 1838** George Bell, Publisher, established in London.
- 1871** John Ruskin sets up George Allen as a publisher.
- 1914** Stanley Unwin buys George Allen and forms George Allen & Unwin.
- 1976** Allen & Unwin, Inc., and Allen & Unwin Australia, Pty. Ltd., formed in Boston and Sydney.
- 1977** Robin Hyman buys George Bell and Bell & Hyman is formed.
- 1986** Allen & Unwin and Bell & Hyman merge interests to form Unwin Hyman.
- 1988** Allen & Unwin, Inc., Boston, becomes Unwin Hyman, Inc.

*Our authors have included such distinguished scholars and writers as:*

Max Weber, Bertrand Russell, Friedrich Nietzsche, Peter Hall, James M. Meade, Charles Kindleberger, Ann Markusen, Alec Nove, Paul Kennedy, James Rosenau, K.J. Holsti, Agnes Heller, Ferenc Feher, Tom Bottomore, Gordon L. Clark, E.N.K. Clarkson, F.K. North.

Our expanding U.S. editorial program now includes: gender studies, Soviet studies, international relations, Latin American studies, and media studies and popular culture.



### **UNWIN HYMAN, INC.**

8 Winchester Place, Winchester, MA 01890  
Toll Free 1-800-547-8889  
In MA and Canada 617-729-0830

# New From Unwin Hyman, Inc.

## **Towards a Radical Democracy**

The Political Economy of the  
Budapest School

*Douglas M. Brown*

March 1988 260pp.

Cloth \$39.95 0-04-330408-7

## **Prosperity and Public Spending**

Transformational Growth and  
the Role of the Government

*Edward Nell*

April 1988 224pp.

Cloth \$34.95 0-04-339044-7

Paper \$16.95 0-04-339045-5

## **The Collected Papers of James Meade**

Volume I: Employment  
and Inflation

*Edited by Susan Howson*

April 1988 656pp.

Cloth \$75.00 0-04-331115-6

## **Floating Exchange Rates**

Theories and Evidence

*Ronald MacDonald*

April 1988 260pp.

Cloth \$45.00 0-04-338134-0

Paper \$24.95 0-04-338135-9

## **Evolutionary Macroeconomics**

*John Foster*

1987 336pp.

Cloth \$49.95 0-04-339041-2

## **Economic Methodology and Freedom to Choose**

*Patrick J. O'Sullivan*

1987 256pp.

Cloth \$50.00 0-04-330375-7

## **Cost-Benefit Analysis in Urban and Regional Planning**

*J.A. Schofield*

1987 256pp.

Cloth \$39.95 0-04-3398145-6

## **Cost Benefit Analysis**

*Fourth Edition*

*E.J. Mishan*

June 1988 460pp.

Paper \$19.95 0-04-445092-3

## **Report of the Task Force on International Debt**

*Rudiger Dornbusch*

April 1988 120pp.

Cloth \$18.95 0-87078-226-6

Paper \$ 9.95 0-87078-225-8

*Priority Press Publication*

## **Taxation by Political Inertia**

Financing the Growth of  
Government in Britain

*Richard Ross & Terence Karran*

1987 224pp.

Cloth \$39.95 0-04-320197-0

Paper \$16.95 0-04-320198-9

For Order Information call:

1-800-547-8889

## ***New from Cambridge University Press***

### **The Boundaries of Economics**

**Gordon C. Winston, Richard F. Teichgraeber, III**

Written by economists and philosophers, these papers examine the themes that complicate the conventional economist's view of the world and thereby provide a notably more complex (and humane) subject than the traditional *homo economicus*.

**Contributors:** Richard F. Teichgraeber, III, Gordon C. Winston, Donald N. McCloskey, John Gray, Michael S. McPherson, Daniel M. Hausman.

Murphy Institute Studies in Political Economy  
\$27.50

### **Income Distribution and the Macroeconomy** **Brian Nolan**

This book looks at the effect that changes in the general condition of the economy, particularly changes in the level of unemployment, have on the distribution of income among persons.

\$39.50

### **The Growth and Efficiency of Government Spending** **Malcolm Levitt and Michael Joyce**

Analyzes the growth and pattern of public spending in Britain since the 1960s and provides alternative estimates for the 1990s. Against this background the authors outline the problems associated with the interpretation of the concept of output in the public services.

National Institute of Economic and Social Research Occasional Paper XLI  
\$39.50

### **Phases of Economic Growth 1850-1973**

**Kondratieff Waves and Kuznets Swings**

**Solomos Solomou**

Examines the evidence for long-term patterns of economic growth. It uses data for Britain, France, Germany, the United States, and the world economy, between 1850 and 1973.

\$44.50

### **Unkept Promises, Unclear Consequences**

**U.S. Economic Policy and the Japanese Response**

**Ryuzo Sato and John A. Rizzo, Editors**

Examines the role of United States macroeconomic policy in the large and growing trade imbalance between Japan and the United States.

**Contributors:** Ryuzo Sato, John Rizzo, Barry Bosworth, Herbert Stein, Roger E. Brinner, Paul A. Samuelson, Lester C. Thurow.

\$24.95

### **The Fall of the Bell System** **A Study in Prices and Politics** **Peter Temin with Louis Galambos**

"...His analysis of the events and the behavior of the various participants, especially the Bell culture's inability to cope with a changing environment, is thorough, unsparing, and objective."—*Science*

\$27.95

## ***New Paperbacks***

### **Foundations in Public Economics**

**David A. Starrett**

Surveying the modern theories of decision making in the public sector, Professor David Starrett synthesizes and interrelates the major theoretical foundations of modern public sector economics.

**\$15.95**

### **Applied Production Analysis**

*A Dual Approach*

**Robert G. Chambers**

This book contains a modern treatment of production economics from a dual perspective, with a special emphasis on recent developments in the field.

**\$18.95**

### **Natural Resource Economics**

*Notes and Problems*

**Jon M. Conrad and  
Colin W. Clark**

This book reviews techniques of dynamic optimization and shows how they can be applied to the management of various resource systems.

**\$14.95**

### **The Theory of Environmental Policy**

*Second Edition*

**William J. Baumol,  
Wallace E. Oates**

A rigorous and comprehensive analysis of the economic theory of environmental policy. The authors present a formal, theoretical treatment of those factors influencing the quality of life.

**\$17.95**

### **The Great Merger Movement in American Business, 1895-1904**

**Naomi Lamoreaux**

"...She combines the insights and methods of both business and economic historians to explain the great merger wave...She has drawn examples from an impressive array of scholarly literature in history, economics, and law, and she has developed careful and clear case studies of the steel and paper industries."

—Reviews in American History

**\$9.95**

**Winner of the Frederick Jackson Turner Award, the New England Historical Associations' Book Award, and the Philip Taft Labor History Award**

### **Out of Work**

*The First Century of Unemployment in Massachusetts*  
**Alexander Keyssar**

"Keyssar has written a highly original and valuable contribution to the study of working-class history, a major work that should influence all future research in the field."—*The New Republic*

Studies in Economic History and Policy:  
The United States in the Twentieth Century

**\$14.95**

---

At bookstores or order from

**Cambridge University Press**

32 East 57th Street, New York, NY 10022.

Cambridge toll-free numbers for orders only:

800-872-7423, outside NY State.

800-227-0247, NY State only.

MasterCard and Visa accepted.



## Long Cycles

*Prosperity and War in the Modern Age*

Joshua S. Goldstein

In this pathbreaking interdisciplinary work, Joshua S. Goldstein tackles the issue of long cycles and their role in the development of the world system. After providing the first comprehensive review of existing literature in this field, Goldstein reports his own empirical findings based on a statistical analysis of over fifty historical time series, constructs a new interpretation of world history in the modern age, and projects cyclical dynamics into the twenty-first century.

"[A] fascinating and truly erudite work."

—Bruce Russett *New in cloth* (\$45.00)

*and paper* (\$19.95)

## The Nonprofit Sector

*A Research Handbook*

edited by Walter W. Powell

A state-of-the-art survey of the political, economic, cultural, legal, sociological, and management aspects of nonprofit organizations in America and abroad.

"This book is likely to be the bible of researchers on the nonprofit sector for the next decade. It is a superb, comprehensive, and thoughtful piece of work." —Stanley N. Katz  
\$45.00

*Now available in paperback*

## What is Political Economy?

*A Study of Social Theory and Underdevelopment*  
Martin Staniland

"Economists and political scientists will find in this work a useful bridge for understanding political economy as practiced by those from the respective disciplines." —Ronnie J. Phillips,  
*Journal of Economic Issues* \$8.95

## Journal of Law, Economics, and Organization

edited by Jerry Mashaw and Oliver Williamson  
One-year subscription (two issues): \$22.00  
individuals, \$30.00 institutions

*Now available in paperback*

## Stabilizing an Unstable Economy

Hyman P. Minsky

A respected economist provides a pathbreaking financial theory of investment to explain the unstable behavior of the American economy and recommends ways to stabilize it under conditions of high employment and non-inflationary prices.

"The year's bid for the most relevant and realistic economic text." —Elliot Janeway,  
*Commonweal* \$14.95

*A Twentieth Century Fund Report*

## Policy, Power, and Order

*The Persistence of Economic Problems in Capitalist States*

Kerry Schott

"Exciting and innovative. [Schott] is a true political economist, who actually uses economic theories to understand nonmarket behavior and who actually makes the distribution of power part of macroeconomic theory."  
—*Politics and Society*

"A stimulating book on economic theories of the state." —Frederick van der Ploeg,  
*The Economic Journal* \$10.95

## Forecasting Political Events

*The Future of Hong Kong*

Bruce Bueno de Mesquita, David Newman, and Alvin Rabushka

What will happen to Hong Kong when Great Britain transfers sovereignty and administrative authority to China in 1997? This book forecasts Hong Kong's future by applying an innovative formal interest group theory of politics.

"Thorough and perceptive." —John Walden,  
*Asian Wall Street Journal* \$12.95



Yale University Press  
Dept. 714  
92A Yale Station  
New Haven, CT 06520

# PROFITABLE READING

New Scholarship on Economics.

## **The Economy of Colonial America** Second Edition

**Edwin J. Perkins**

A comprehensive, up-to-date synthesis of the scholarly literature on the economy of Colonial America. This second edition reflects the enormous amount of new research on the colonial period that has been done in the past decade, most notably on the role of women in the colonial economy.  
264 pp., maps, \$13.00 pa, \$30.00 cl

## **New in the POLITICAL ECONOMY OF INTERNATIONAL CHANGE Series** **John Gerard Ruggie, General Editor**

*Winner of the Edwin W. Rickert Award*

### **Managing International Markets** Developing Countries and the Commodity Trade Regime

**Jock A. Finlayson and Mark W. Zacher**

An analysis of one of the most important issues in North-South negotiations in recent decades: the intergovernmental regulation of commodity markets.  
352 pp., tables, \$40.00

### **Sanctity Versus Sovereignty**

The United States and the Nationalization  
of Natural Resource Investments

**Kenneth A. Rodman**

Traces how American government and corporate officials have resisted—and adapted to—third world economic nationalism.  
448 pp., tables, \$45.00

### **Making Sense of Europe**

**Christopher Tugendhat**

“A brilliantly informative book that includes both an analysis of what exists and a set of practical guidelines for further progress.”

—Fritz Stern, *Foreign Affairs*

“A major treatise about the whole European condition...” —*The Observer*

240 pp., \$25.00

### **Rural Poverty in South Asia**

**Edited by T.N. Srinivasan  
and Pranab K. Bardhan**

An analytical and quantitative study of patterns, trends, and policies of poverty alleviation in rural South Asia. The authors explain the variations in poverty over time and area, and among socio-economic groups.

608 pp., tables, illus., \$50.00

## **Restructuring the Automobile Industry**

A Study of Firms and States  
in Modern Capitalism

**Dennis P. Quinn, Jr.**

A sophisticated, detailed analysis of the automobile industry that offers a lucid argument concerning state-industry relations.

*Columbia Studies in Business, Government,  
and Society*

*Eli M. Noam, General Editor*

395 pp., tables, graphs, \$40.00

## **Custom and Contract**

Household, Government, and the  
Economy in Colonial Pennsylvania

**Mary M. Schweitzer**

Original and imaginative, *Custom and Contract* examines the relationship between local governments and individual households in Pennsylvania during the first half of the eighteenth century.

228 pp., tables, graphs, \$32.00



**PINTER PUBLISHERS**

distributed in the U.S. and Canada

## **Lloyds Bank Annual Review**

Privatization and Ownership

**Edited by Christopher Johnson**

An annual thematic volume that draws on the resources of Lloyds Bank Review and features original articles by distinguished economists.  
200 pp., \$30.00

## **Long-Run Economics**

An Evolutionary Approach to  
Economic Growth

**Norman Clark and Caletous Juma**

Case studies illustrate that technological change is a key characteristic of socioeconomic transformation and that economic systems are propelled through time by the innovations originating within the science and technology system itself.

230 pp., \$30.00

To order, send check or money order,  
including \$3.00 for postage and handling, to:



**COLUMBIA  
UNIVERSITY PRESS**

Dept. JN 136 South Broadway,  
Irvington, NY 10533

## New Titles from Harvard University Press

### **FREE TO LOSE**

An Introduction to Marxist  
Economic Philosophy

**John E. Roemer**

"Roemer's reconstruction of the concept of exploration and his elaboration of its relationship to class constitute the most important theoretical innovations on these problems in contemporary Marxism."—Erik Olin Wright, University of California at Berkeley

\$22.50 cl.; \$8.95 p.

### **THE NONPROFIT ECONOMY**

**Burton A. Weisbrod**

"A major contribution to an important and neglected subject... thoughtfully analytical, well documented, and clearly formulated. It will be valuable for years to come."—David A. Hamburg, M.D., President, Carnegie Corporation of New York

\$22.95

### **THE JAPANESE TODAY**

Change and Continuity

**Edwin O. Reischauer**

The foremost interpreter of Japanese history and culture brings us an incomparable description of Japan today in all its complexity and uniqueness — the history, politics, economy, and international position of this fascinating island nation.

*Belknap*

\$25.00

### **THE CONQUEST OF THE MICROCHIP**

**Hans Queisser**

This fascinating insider's view of the birth of the microelectronics industry provides a unique perspective on an era of new knowledge that has resulted not only in the restructuring of science, technology, and industry, but also in major rearrangements of political and economic power.

\$24.95

*Now available in paperback*

### **THE HEALTH ECONOMY**

**Victor R. Fuchs**

"[Fuchs] is equally adept in probing the psychological motives of individual patients and physicians and in projecting how changes in economic incentives will affect the behavior of vast groups of doctors, hospitals, and insurers."

—*Washington Post*

\$12.95 paper

### **UNTANGLING THE INCOME TAX**

**David F. Bradford**

"Truly first-rate . . . An excellent introduction to many of the major topics in modern public finance and, as such, will make an excellent supplementary text for courses in public finance."

—*Journal of Economic Literature*

\$14.95 paper

**H**arvard University Press  
79 Garden St. Cambridge, MA 02138

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## **COMPARATIVE PATTERNS OF ECONOMIC DEVELOPMENT, 1850-1914**

Cynthia Taft Morris and Irma Adelman

The result of a twenty-year investigation, this book provides an empirical analysis of the dynamics of economic and institutional change in twenty-three countries. The authors marshal an enormous amount of quantitative data and apply an innovative methodology to explore the reasons for success and failure in economic development.

*The Johns Hopkins Studies in Development*

Vernon W. Ruttan and T. Paul Schultz, Consulting Editors

\$39.50

## **AGRICULTURAL PRICE POLICY FOR DEVELOPING COUNTRIES**

edited by John W. Mellor and Raisuddin Ahmed

Distinguished specialists here examine agricultural price policy in the broader context of the technological and institutional changes that "are the essence of development." The book recommends pragmatic approaches for managing price fluctuations and exchange rates, relating domestic to international prices, and balancing the needs of producers and consumers in a coherent and consistent strategy of economic development.

*Published in cooperation with the International Food Policy Research Institute*

\$35.00 hardcover

## **PROJECT MONITORING AND EVALUATION IN AGRICULTURE**

Dennis J. Casley and Krishna Kumar

Monitoring the implementation of projects and evaluating their achievements are vital parts of the project cycle. In this volume the authors provide a wealth of new examples to explain in detail how to monitor and evaluate agricultural and rural development projects.

*Published for the World Bank*

\$20.00 hardcover    \$12.95 paperback

## **THE COLLECTION, ANALYSIS, AND USE OF MONITORING AND EVALUATION DATA**

Dennis J. Casley and Krishna Kumar

This volume provides practical methods of collecting and analyzing data for monitoring and evaluating agricultural projects. Because of the limited resources of many development projects, the authors have selected methods that are both simple and inexpensive. Together with its companion volume, *Project Monitoring and Evaluation in Agriculture*, this book will be useful for those who design and implement these systems and as a text for regional and national training programs.

*Published for the World Bank*

\$22.50 hardcover    \$14.50 paperback



**THE JOHNS HOPKINS UNIVERSITY PRESS**

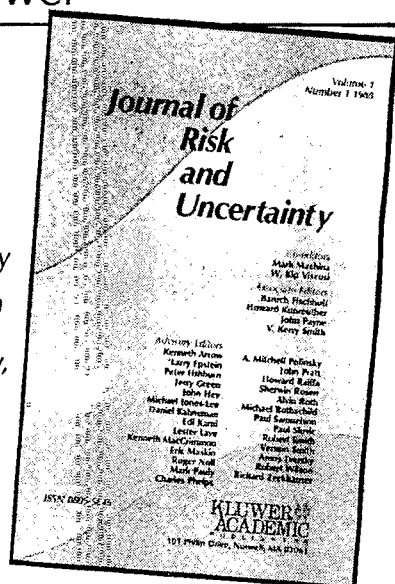
701 West 40th Street, Suite 275, Baltimore, Maryland 21211

An exciting new journal from Kluwer

# ***The Journal of Risk and Uncertainty***

Co-Editors: **Mark Machina**, University of California, San Diego, and **W. Kip Viscusi**, Northwestern University

Associate Editors: **Baruch Fischhoff**, Carnegie Mellon University, **Howard Kunreuther**, Wharton School, University of Pennsylvania, **John Payne**, Duke University, and **V. Kerry Smith**, North Carolina State University



## *Papers included in Volume 1, 1988*

*Risk Ambiguity and Insurance*  
by H. Kunreuther and R. Hogarth

*Aversion to One Risk in the Presence of Others*  
by J.W. Pratt

*Status Quo Bias in Individual Decision Making*  
by R. Zeckhauser and W. Samuelson

*Risk Aversion in Bargaining: An Experimental Study*  
by A. Roth, K. Murnighan and F. Schoumaker

**The Journal of Risk and Uncertainty** welcomes original manuscripts, both theoretical and empirical, dealing with the analysis of risk-bearing behavior and decision making under uncertainty. The topics covered in the journal include, but are not limited to:

- decision theory and the economics of uncertainty
- psychological models of choice under uncertainty
- risk and public policy
- experimental investigations of behavior under uncertainty
- empirical studies of real work risk-taking behavior

An important aim of the **JRU** is to encourage interdisciplinary communication and interaction between researchers in the area of risk and uncertainty.

**Subscription Information** Rates per volume; 1988 (4 issues) including postage and handling  
Individual: \$45.00 Institutional: \$107.00 ISSN 0895-5646

Information for Authors, sample copy requests,  
and subscriptions should be addressed to:

**KLUWER**   
**ACADEMIC**  
PUBLISHERS

101 Philip Drive • Norwell, MA 02061

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# AMERICAN ECONOMIC ASSOCIATION

## 1988 ANNUAL MEMBERSHIP RATES

### Membership includes:

— a subscription to *The American Economic Review* (quarterly) plus *Papers and Proceedings*, the *Journal of Economic Literature* (quarterly) and the *Journal of Economic Perspectives* (quarterly).

- Regular members with annual incomes of \$30,000 or less ..... \$38.50
- Regular members with annual incomes above \$30,000 but no more than \$40,000 ..... \$46.20
- Regular members with annual incomes above \$40,000 ..... \$53.90
- Junior members (available to registered students for three years only).

Student status must be certified by your major professor or school registrar ..... \$19.25

- In Countries other than the U.S.A., Add \$16.00 to cover postage.
- Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association) ..... \$7.70

Please enter my subscription for the following period:

☐ Jan.-Dec.      ☐ April-March      ☐ July-June      ☐ Oct.-Sept.

First Name and Initial	Last Name	Suffix
Address Line 1		
Address Line 2		
City		
State or Country	Zip/Postal Code	

**MAJOR FIELDS (TWO ONLY)**  
LIST FIELDS WITH WHICH  
YOU CURRENTLY IDENTIFY.  
SELECT FIELD CODE FROM JEL,  
"Classification System  
for Books."

--	--

Please type or print information above. Please pay with a check or money order payable in United States Dollars. Canadian and foreign payments must be in the form of a draft or check drawn on a United States bank payable in United States Dollars. Please note: It is the policy of the Association, not to refund membership payments.

Endorsed by (AEA member) \_\_\_\_\_

### Below for Junior Members Only

I certify that the person named above is enrolled as a student at \_\_\_\_\_

\_\_\_\_\_  
Authorized Signature

PLEASE SEND WITH PAYMENT TO:

**AMERICAN ECONOMIC ASSOCIATION**  
1313 21ST AVENUE SOUTH, SUITE 809  
NASHVILLE, TENNESSEE 37212-2786  
U.S.A.

# AEA sponsored Group Life Insurance for you and your family— at attractive rates!

The AEA Group Life Insurance Plan can help provide valuable supplementary protection—at attractive rates—for eligible members and their dependents.

Because AEA participates in a large Insurance Trust which includes other scientific and technical organizations, the low cost may be even further reduced by premium credits. In the past nine years, insured members received credits on their April 1 semiannual payment notices averaging 40% of their annual premium contributions. (These credits are based on the amount paid during the previous policy year ending September 30.) Of course future premium credits, and their amounts, cannot be promised or guaranteed.

Now may be a good time for you to re-evaluate your present coverage and look into AEA Life Insurance. Just fill out and return the coupon for more details at no obligation.

<b>Administrator, AEA Group Insurance Program</b>		J-3
1255 23rd Street, N.W. Washington, D.C. 20037		
Please send me more information about the AEA Life Insurance Plan.		
Name _____	Age _____	
Address _____		
City _____	State _____	Zip _____

Or—call today Toll-Free 800-424-9883  
(Washington, DC area, call 296-8030)

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*



# NEW FOR 1988

## Titles in Economics from South-Western

**New**

### **ECONOMICS: A Contemporary Introduction**

William A. McEachern, University of Connecticut, Storrs

This engaging new text provides a fresh, contemporary approach to economic principles. It employs numerous examples from everyday life to illustrate fundamental economic concepts as well as recent developments in the field. The supplementary package is the most comprehensive available.

**New**

### **MACROECONOMICS: Theory and Policy**

Steven M. Sheffrin, University of California, Davis

David A. Wilton, University of Waterloo

David M. Prescott, University of Guelph

A lively discussion of policy and its applications sets this text apart. Theory is developed based on the IS-LM/PEP curve framework, and coverage of current issues such as hysteresis, destabilizing price flexibility, exchange rate overshooting, and deficits brings students up to date on the latest topics of economic study.

**Just  
Revised**

### **MACROECONOMICS: Intermediate Theory and Policy, Second Edition**

William J. Boyes, Arizona State University

This student-oriented text includes a thorough discussion of theory, as well as basic macro policy issues and background information. International issues such as exchange rates and the open economy are covered in detail.

**Just  
Revised**

### **MANAGERIAL ECONOMICS: Analysis, Problems, Cases, Third Edition**

Lila Truett and Dale Truett

University of Texas, San Antonio

The central focus of this text is the relevance of profit-maximizing principles for today's business manager. The first five chapters are devoted to an explanation of all the tools and decision criteria (revenue, demand, production, cost, profit-maximization) that your students need to analyze problems later in the text. The expanded supplementary package now includes The Decision Assistant, an IBM microcomputer applications package containing tools for economic analysis.

**For more information on these or other current economics texts, contact the South-Western representative in your area.**

**SOUTH-WESTERN**  
COLLEGE DIVISION

5101 Madison Road • Cincinnati, OH 45227





---

# Work, Games, and Government . . .

## New

---

### **THE WORKPLACE WITHIN**

Psychodynamics of Organizational Life  
**Larry Hirschhorn**

Facing a postindustrial market, groups at work become increasingly irrational. In this revealing study, Larry Hirschhorn brings his considerable consulting experience to bear on developing a theory of social defenses—thoughtlessness, ritualization, irrational behavior—that are becoming increasingly costly to society. Included are recommendations for dealing with these destructive new organizational behaviors.  
\$16.95

### **A GENERAL THEORY OF EQUILIBRIUM SELECTION IN GAMES**

**John C. Harsanyi and Reinhard Selten**  
Foreword by Robert Aumann

"The book provides for the first time a heroic and thorough attempt to suggest a very general selection theory. . . . A successful theory of this type may change all of economic theory. The book is likely to be controversial among game-theoreticians and economists, but none would be able to dismiss its importance."—Ariel Rubinstein, Hebrew University  
July \$32.50

### **THE MASSACHUSETTS MIRACLE**

High Technology and Economic  
Revitalization  
*edited by David R. Lampe*

*The Massachusetts Miracle presents a col-*

lection of highly readable opinion pieces and policy statements written by local economists and policymakers over the last decade and a half describing Massachusetts's dramatic transformation into a robust high-tech economy.  
\$16.95

### **MARKETS OR GOVERNMENTS**

Choosing between Imperfect Alternatives  
**Charles Wolf, Jr.**

Economic studies historically have either extolled the virtues of perfect markets or decried the market's shortcomings, proposing that governments correct market failure. This book proposes as a counterweight to these views a theory of nonmarket failure, and examines in great detail the shortcomings of government efforts to replace or to regulate markets.  
\$18.95

*New in Paperback*

### **TWO REVOLUTIONS IN ECONOMIC POLICY**

The First Economic Reports of Presidents  
Kennedy and Reagan

*edited by James Tobin and  
Murray Weidenbaum*

*with introductions by James Tobin and Robert Solow  
and by William Niskanen, William Poole, and  
Murray Weidenbaum*

The juxtaposition of Kennedy to Reagan approaches to economic problems is particularly instructive in that they express the two major—and quite different—approaches of macroeconomic policy in the past three decades. This book reproduces the original economic reports together with valuable commentary by leading economists.  
July \$13.95 paper (\$30.00 cloth)

---

## **The MIT Press**

55 Hayward Street, Cambridge, MA 02142

---



# MANHATTAN COLLEGE SCHOOL OF BUSINESS

## FACULTY OPENINGS IN ECONOMICS AND FINANCE

Medium-sized, independent, private college seeks faculty in Economics and Finance for graduate and undergraduate programs. Ph.D/D.B.A., teaching experience, publications for upper ranks. A.B.D. considered for lower ranks. Salary competitive in category 1+ of AAUP ratings. Attractive fringe benefits.

Send resume to: Dr. Faraj Abdulahad, Dean, School of Business, MANHATTAN COLLEGE, Riverdale, NY 10471. An EO/AA Employer

## ECONOMETRIC SOFTWARE FROM TSP INTL

### *Now available for PCs and mainframe computers*

**TSP Version 4.1:** with Probit, Logit, Tobit, sample selection, general maximum likelihood estimation, and robust standard errors for all procedures. A complete programming language for econometricians and data analysts in use at over 1000 sites worldwide, it includes regression, nonlinear simultaneous equation estimation, time series models, and many other features.

The PC version is interactive and identical to the mainframe version. Databanks and the saving and restoring of workspaces are now supported. PC TSP requires 512K RAM, a math chip (8087 or 80287), hard disk recommended.

### *For mainframes only*

**RATS Version 2.0:** identical to the popular PC version. We are the authorized distributor of the mainframe version.

For more info, write or call (415) 326-1927  
TSP International • PO Box 61015 • Palo Alto, CA 94306



# Our 1988 List is Dynamite!



## **ECONOMICS**

**ECONOMICS, 2/e**

**INTRODUCTION TO MICROECONOMICS, 2/e**

**INTRODUCTION TO MACROECONOMICS, 2/e**

**Stanley Fischer, Rudiger Dornbusch, and  
Richard Schmalensee**, all of the  
Massachusetts Institute of Technology

**ECONOMETRICS, 2/e**

**Damodar N. Gujarati**,  
Bernard Baruch College, C.U.N.Y.

**INTRODUCTION TO MACROECONOMICS**

**1987-1988: Readings on  
Contemporary Issues**

**Peter D. McClelland**, Cornell University

## **FINANCE**

**PRINCIPLES OF CORPORATE FINANCE, 3/e**

**Richard A. Brealey**, London Graduate  
School of Business Studies

**Stewart C. Myers**, Massachusetts Institute  
of Technology

**INTRODUCTION TO FINANCIAL  
MANAGEMENT, 5/e**

**Lawrence D. Schall** and  
**Charles W. Haley**, both of the  
University of Washington

**INTRODUCTION TO CORPORATE FINANCE**

**Terry S. Maness**, Baylor University

**STRATEGY FOR PERSONAL FINANCE, 4/e**

**Larry R. Lang**, The University of Wisconsin

**MANAGEMENT OF INVESTMENTS, 2/e**

**Jack Clark Francis**, Baruch College, C.U.N.Y.

**READINGS AND CASES IN  
CORPORATE FINANCE**

**Stephen Archer**, Willamette University

**Halbert S. Kerr**, Washington State University

**GUIDE to IFPS/Personal**

**Paul Gray**, Claremont Graduate School of Business

*To receive an examination copy, please write:*



**McGraw-Hill Book Company**

College Division P.O. Box 443 Hightstown, New Jersey 08520

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers.*

# New Economics titles from

## Economics

### *Between Predictive Science and Moral Philosophy*

James M. Buchanan

Compiled by Robert D. Tollison  
and Viktor J. Vanberg

Nobel Laureate James M. Buchanan played a key role in the development of "theoretical institutional economics"—and was awarded the 1986 Nobel Prize in economic science for his contributions to a theory of political economy as well as his leadership of the public choice movement.

These twenty-six papers, which form the core of the author's work, span various subfields of economics from public finance to methodology, and in each paper the constitutional economics paradigm—viewed by the author as a modern revival of classical political economy—is modified, extended, and applied to particular issues. 432 pp. 28 line drawings. \$48.50

### *Spatial Price Theory of Imperfect Competition*

Hiroshi Ohta

Economic space has long been a neglected element in orthodox economic theory, one thought to complicate the issue unnecessarily. But the theoretical implications of assuming away spatial elements may be especially significant for pricing practices and hence for competition.

Hiroshi Ohta shows why and in what ways economic space is needed to reform orthodox price theory. Negating the classical paradigm of perfect competition, he calls for a spatial price theory of imperfect competition. Among his findings in spatial microeconomic theory are that unlimited entry of new firms into the market may not lower consumer prices and that increased labor productivity in a spatial economy may actually lower real wages. Ohta's work is particularly valuable in understanding a decade of advances in spatial price theory and exploring new theories of competition. 232 pp. 51 line drawings. \$34.50

NEW IN PAPERBACK

### *Mercantilism as a Rent-Seeking Society Economic Regulation in Historic Perspective*

Robert B. Ekelund, Jr.  
Robert D. Tollison

Using positive-economics principles, Robert B. Ekelund and Robert D. Tollison show how rent seeking provided first the impetus for European mercantilism and later the reasons for its demise in England and entrenchment in France. The balance-of-trade objective is revealed as the by-product of self-interested parties' seeking of rents.

In addition to questioning the causes and results of economic regulation, the authors raise issues in the methodology of economic history that will particularly appeal to public choice theorists, political economists, and economic policy-makers.

"... balanced and informative, and well worth reading."—*History of Political Economy*. 184 pp. \$12.95 paper

### *Governmental Controls and the Free Market*

#### *The U.S. Economy in the 1970's*

Edited by Svetozar Pejovich

The contributors to this book, including Armen Alchian, William A. Niskanen, James M. Buchanan, and Gordon Tullock, are concerned with dangers to the institutions of private property, with profit incentives, and with social interaction guided by principles of self-interest and open-market competition.

Topics include public attitudes toward free enterprise; inflation; unemployment; the political process; social security; government transfer spending; the efficiency of the modern corporation; and worker alienation and the structure of the firm.

"... remarkable book."—*Southern Economic Journal*. 240 pp. \$8.95 paper

# Texas A&M University Press

DRAWER C COLLEGE STATION, TEXAS 77843-4354 409-845-1436

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

## Computer Access to Articles in the JEL Subject Index

Online computer access to the *JEL* and *Index of Economic Articles* database of journal articles is currently available through DIALOG Information Retrieval Service. DIALOG file 139 (*Economic Literature Index*) contains complete bibliographic citations to articles from the nearly 300 journals listed in the quarterly *JEL* issues from 1969 through the current issue. The abstracts published in *JEL* since June 1984 are also available as part of the full bibliographic record. The *Economic Literature Index* also includes citations to articles in the 1979 and 1980 collective volumes (collected papers, proceedings, etc.) for the *Index* database; other years will be added as soon as completed. The file may be searched using free-text searching techniques or author, journal, title, geographic area, date, and other descriptors, including descriptor codes based on the *Index's* four-digit subject classification numbers. (For a complete description of the *Economic Literature Index* with search examples and suggestions for searching techniques, see the article "Online Information Retrieval for Economists—The Economic Literature Index," in the December 1985 issue of the *Journal of Economic Literature*.)

### *Access Options:*

- **DIALOG** offers a variety of contract choices, including the option (for a low annual fee) to pay for only what you use. Most university libraries already subscribe to DIALOG. For information on the DIALOG service, contact your librarian or write to or call: DIALOG Information Services, Inc., Marketing Department, 3460 Hillview Avenue, Palo Alto, California 94304 (800-3-DIALOG or 800-334-2564).
- **Knowledge Index**, a DIALOG service available after 6 p.m. and on weekends, may be accessed at the low rate of \$24/hour, charged to a major credit card. A one time start-up fee of \$35.00 buys 2 hours free time during the first month after log-on. Call 800-3-DIALOG for information.
- **EasyNet**, a gateway service, provides menus to guide the untrained user through database searches in DIALOG and other databases. For information, call 1-800-841-9553 or dial up EasyNet on your terminal (1-800-EASYNET) and pay for your search by credit card.

### *Classroom Instruction:*

- **DIALOG's Classroom Instruction Program**, available at a special rate of \$15/connect hour to academic institutions for supervised instruction, permits teachers to incorporate online bibliographic searching in their courses. For information, contact DIALOG or your librarian.

*New from the Consortium of Social Science Associations....*

The first comprehensive

## **Guide to Federal Funding for Social Scientists**

Prepared by the Consortium of Social Science Associations (COSSA), a Washington advocacy group serving the major professional societies in the social and behavioral sciences. Susan D. Quarles, editor.

The federal government is a major supporter of research in the social and behavioral sciences, but until now, no single, multidisciplinary directory has been available to guide researchers through the complexities of government funding in these fields.

COSSA's inclusive new *Guide to Federal Funding* describes over 300 federal programs in impressive detail, including funding priorities, application guidelines, and examples of funded research. Introductory essays describe the organization of social science funding and offer inside views of federal funding practices and contract research.

For anyone who needs to know the ins-and-outs of government funding in the social sciences and related fields, COSSA's *Guide* will be an essential new resource.

**Published by the Russell Sage Foundation**

512 pages ISBN 0-87154-699-X Paperback

**Mail orders to:**

Consortium of Social Science  
Associations  
1625 I St, NW, Suite 911  
Washington, D.C. 20006

Orders will be filled when books are available,  
in June. Please allow 3-4 weeks for delivery.

Please send \_\_\_\_\_ copies of COSSA's *Guide  
to Federal Funding for Social Scientists*  
(81-0699X) at the following price:

- ☐ \$24.95 (for libraries and institutions)  
☐ **\$14.95 (special price for  
members of AEA)**  
☐ \$19.95 (for other individuals)

Enclosed is a check, money order, or purchase  
order in the amount of \$ \_\_\_\_\_. Publisher  
pays postage on prepaid orders; New York  
residents; please add sales tax.

\_\_\_\_\_  
Name (please print)

\_\_\_\_\_  
Address  
\_\_\_\_\_

# Journal of International Economic Integration

**Solicits Papers to Compete for the  
Annual Daeyang Prize in Economics (\$7,000)  
and Welcomes Subscriptions by Interested Parties**

## **Current issues include**

**Bela Balassa**, *Japanese Trade Policies Towards Developing Countries.*

**Basant K. Kapur**, *Open-Economy Response to a Terms of Trade Shock in a Growth Context.*

**Norman C. Miller**, *A General Approach to the Balance of Payments and Exchange Rates.*

**Leonard F.S. Wang**, *Product Market Imperfections and Customs Unions theory.*

**Gene M. Grossman**, *The Employment and Wage Effects of Import Competition in the United States.*

**Wilfred J. Ethier and Ronald D. Fischer**, *The New Protectionism*

**Joshua Aizenman**, *Inflation, Tariffs and Tax Enforcement Costs*

**Chung H. Lee and Seiji Naya**, *The Internationalization of U.S. Service Industries and Its Implications for Developing Countries.*

The Journal of International Economic Integration is published biannually (Spring and Autumn) by the Institute for International Economics, King Sejong University, Seoul, Korea.

The purpose of the Journal is to support and encourage research in the area of international trade, international finance and other related economic issues that include general professional interest in international economic affairs. Welcoming both theoretical and empirical analyses in international economics, the Journal is strongly interested in the issues of the international economic cooperation.

- The Journal welcomes unsolicited manuscripts, which will be considered for publication by the Editorial Board.
- From papers selected for publication, the Prize committee will choose the best manuscript(s) to receive the \$7,000 Daeyang Prize. The winner of the prize is announced in the Spring issue every year.
- The manuscripts should be accompanied by an abstract of no more than 100 words and a brief curriculum vitae containing the author's academic career. All submissions should be typewritten, double-spaced, in English with footnotes, references, figures, tables and any other illustrative material on separate sheets.
- Three copies of the manuscript and all accompanying material should be submitted to the following address by October 31, 1988 for consideration for 1989 publication.
- For subscriptions to the Journal (\$20 per year for individuals, \$30 per year for institutions), send a check or money order payable to King Sejong University to the following address.

**Institute for International Economics  
King Sejong University  
Seongdong-Ku, Seoul, Korea**

O C D E



O E C D

### **The Costs of Restricting Imports: The Automobile Industry.**

Presents the findings of studies by independent analysts of the effects of restrictions on imports and sales of foreign cars in four OECD countries: the United States, Canada, France, and the United Kingdom. It also demonstrates the usefulness of a checklist devised by the OECD and recommended to governments of Member countries in 1985 to help them assess the impact of proposed and existing regulations on trade in all products.

24-87-06-1, January 1988, 173 pages, ISBN 92-64-13037-3, \$18.00

### **National Accounts Volume 1: Main Aggregates 1960-1986.**

The 1988 edition of one of OECD's most asked-for publications. Contains graphs for each OECD country showing GDP, Private and Government Final Consumption Expenditure, and Gross Fixed Capital Formation; tables for each country showing the main aggregates in national currencies; "growth triangles" showing percent changes for the main comparative tables in U.S. dollars and in Purchasing Power Parities.

30-88-01-3, February 1988, 151 pages, ISBN 92-64-03017-4, \$27.00

### **Purchasing Power Parities and Real Expenditures, 1985.**

Final and detailed national accounts information, broken out by sector and presented in purchasing power parity form, in both the SNA and ICP classification systems.

30-87-06-3, January 1988, 63 pages, ISBN 92-64-03018-2, \$15.50

### **Industrial Structure Statistics, 1985.**

Statistics on value added, production, employment, investment, and wages and salaries in over 65 industries in OECD countries covering the years 1982-1985.

70-87-03-3, January 1988, 145 pages, ISBN 92-64-03019-0, \$19.80

### **External Debt Statistics.**

This report, containing statistics on the volume and composition of the external debt of 155 countries in 1985 and 1986, covers more countries than any other publication of its kind. The way in which the figures were compiled enables the reader to make more comparisons than is usually possible. The report also includes estimates of the amortisation payments each country was due to make on its long-term debt in 1987. Full technical explanations are provided.

43-87-05-1, January 1988, 29 pages, ISBN 92-64-13040-3, \$11.00

*To order, send your check or money order to:*

#### **OECD Publications and Information Center**

2001 L Street, NW, Washington, DC 20036-4095

Telephone: (202) 785-6323

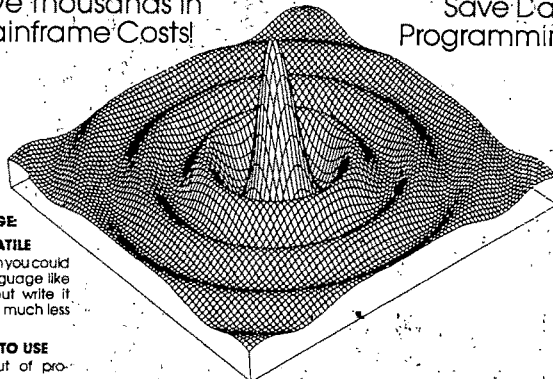


# GAUSS

## The New Standard for Scientific and Statistical Computation.

Save Thousands in  
Mainframe Costs!

Save Days in  
Programming Time!



### SOME FEATURES OF THE GAUSS PROGRAMMING LANGUAGE

#### FULL-FEATURED AND VERSATILE

Write essentially any program you could write in a conventional language like FORTRAN, PASCAL, or C, but write it faster, more easily, and with much less code.

#### EASY TO LEARN AND EASY TO USE

GAUSS takes the work out of programming.

#### EXTREMELY FAST

GAUSS provides the fastest computation, and the fastest I/O, of any program available for PC's, by far. On a plain PC, invert a 50x50 matrix in 14.66 seconds, compute the sums and means of variables in a data set with 10,000 observations, 50 variables in under 1 minute.

#### POWERFUL

GAUSS provides the basic tools of modern applied mathematics, and makes it easy to apply those tools.

#### NUMERICALLY ACCURATE

GAUSS uses state of the art numerical algorithms (LINPACK, EISPACK), and in addition takes optimal advantage of the extended precision of the 8087 numeric coprocessor.

#### COMMENTS FROM GAUSS USERS

"I used to use FORTRAN and PASCAL for languages, TSP and Minitab for statistics, MATLAB for math, and NAG and IMSL for FORTRAN subroutines. Now I just use GAUSS."

Dr. Cheon-Geol Moon  
Rutgers University

"GAUSS is the most beautifully designed software I have ever seen."  
Professor Warren Sanderson  
SUNY Stony Brook

"Having the power of a mainframe on your desktop is more than just a convenience, it will certainly open new avenues in micro computing."  
Frank Siba — London School of Economics

### OTHER FEATURES OF THE GAUSS PROGRAMMING LANGUAGE INCLUDE:

- full-screen editor, screen, printer, file, and keyboard I/O
- specialized functions for statistics and data handling
- state-of-the-art random number generators
- complex arithmetic; polynomial operators; trig functions
- probability density and cumulative distribution functions
- sequence functions, functions for recursive series
- arithmetic, relational, and logical operators
- numerical integration and differentiation
- LINPACK, EISPACK, and related algorithms, including LU, Cholesky, QR, and SVD decompositions; general and positive definite inverses; pseudo inverse; general and positive definite equation solutions; real general and symmetric eigenvalues and eigenvectors
- Coded mostly in Assembler, core program smaller than 200K

## The GAUSS Mathematical & Statistical System

- **DATABASE MANAGEMENT** (enter, convert, edit, sort, merge)
- **STATISTICS** (means, frequencies, crosstabs, regression, non-parametrics, general max. likelihood, non-linear least squares, simultaneous equations, logit, probit, loglinear models, & more)
- **PUBLICATION QUALITY GRAPHICS** (2D & 3D; color, hidden line removal, zoom, pan; up to 4096 x 3120 resolution; produce Tektronix format files; output to most screen drivers, plotters, printers)
- **PLUS:**
  - SIMULATION
  - TIME SERIES/SIGNAL PROCESSING
  - LINEAR PROGRAMMING
  - NON-LINEAR OPTIMIZATION
  - NON-LINEAR EQUATION SOLUTION
  - INTERACTIVE MATRIX PROGRAMMING
  - LARGE-SCALE MODULAR PROGRAMMING
  - ADD YOUR OWN COMMANDS
  - LINK FORTRAN, C, ASSEMBLER SUBROUTINES

Call or Write:

**APTECH  
SYSTEMS, INC.**

1914 N. 34th St., Suite 301  
Seattle, WA 98103  
(206) 547-1733

# THE GAUSS

## MATHEMATICAL AND STATISTICAL SYSTEM

for IBM PC-XT-AT-System/2 and Compatibles  
written by Lee E. Edleisen and Samuel D. Jones

Buy the GAUSS Programming Language by itself or as part of the GAUSS Mathematical and Statistical System, which includes 2D & 3D graphics plus over 200 applications programs written in the GAUSS Programming Language for doing a variety of mathematical, statistical, and scientific tasks. Full source code is provided for these programs.

### 30 DAY MONEY-BACK GUARANTEE

The GAUSS Mathematical and Statistical System .....	\$350
The GAUSS Programming Language (alone) .....	\$200
Shipping/handling, continental USA, 2nd Day Air .....	\$8.50
Shipping/handling, continental USA, Ground .....	\$5.00

GAUSS requires 320K (512K required for high resolution graphics) DOS 2.10+, and a math coprocessor.

NOT COPY PROTECTED

Please mention this publication when responding to this ad.

AS-12/87